Whispering - The hidden side of auditory communication

Fruehholz, Sascha; Trost, Johanna Wiebke; Grandjean, Didier Maurice

How to cite

FRUEHHOLZ, Sascha, TROST, Johanna Wiebke, GRANDJEAN, Didier Maurice. Whispering - The hidden side of auditory communication. In: NeuroImage, 2016, vol. 142, p. 602–612. doi: 10.1016/j.neuroimage.2016.08.023

This publication URL: https://archive-ouverte.unige.ch/unige:95990
Publication DOI: 10.1016/j.neuroimage.2016.08.023

NeuroImage 142 (2016) 602–612

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Whispering - The hidden side of auditory communication

Sascha Frühholz [a,b,c,d,*], Wiebke Trost [d], Didier Grandjean [d,e]

[a] Department of Psychology, University of Zurich, Zurich 8050, Switzerland
[b] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich 8057, Switzerland
[c] Center for Integrative Human Physiology (ZIHP), University of Zurich, 8057, Switzerland
[d] Swiss Center for Affective Sciences, University of Geneva, Geneva 1202, Switzerland
[e] Department of Psychology, University of Geneva, Geneva 1205, Switzerland

ARTICLE INFO

Article history:
Received 28 May 2016
Accepted 11 August 2016
Available online 12 August 2016

Keywords:
Voice
Whispering
Emotion
Auditory system
Neural network
fMRI

ABSTRACT

Whispering is a unique expression mode that is specific to auditory communication. Individuals switch their vocalization mode to whispering especially when affected by inner emotions in certain social contexts, such as in intimate relationships or intimidating social interactions. Although this context-dependent whispering is adaptive, whispered voices are acoustically far less rich than phonated voices and thus impose higher hearing and neural auditory decoding demands for recognizing their socio-affective value by listeners. The neural dynamics underlying this recognition especially from whispered voices are largely unknown. Here we show that whispered voices in humans are considerably impoverished as quantified by an entropy measure of spectral acoustic information, and this missing information needs large-scale neural compensation in terms of auditory and cognitive processing. Notably, recognizing the socio-affective information from voices was slightly more difficult from whispered voices, probably based on missing tonal information. While phonated voices elicited extended activity in auditory regions for decoding of relevant tonal and time information and the valence of voices, whispered voices elicited activity in a complex auditory-frontal brain network. Our data suggest that a large-scale multidirectional brain network compensates for the impoverished sound quality of socially meaningful environmental signals to support their accurate recognition and valence attribution.

Introduction

Francis Ford Coppola's film “The Godfather” opens with the famous whispering scene where Bonasera, after some introductory conversation, fearfully whispers a displeasing request to Don Corleone. The scene is a classic example of how individuals switch their vocalization mode to whispering when deeply moved by their emotions in certain social contexts. Whispering has both biological and social functions. Socially, whispering is used to confine communication to listeners in immediate proximity to signal closeness and secrecy, which supports in-group-cohesion (Cirillo and Todt, 2002) and prevents eavesdropping (Morrison and Reiss, 2013), respectively. Sometimes whispering is also used to completely hide one's own affective states and feelings from being recognized by others. Its biologic functions are demonstrated in some animals, which, for example, switch to vocal whispering when confronted with a predator (Morrison and Reiss, 2013) or to initiate sexual behavior (Ladich, 2007). This unique mode of vocal communication is specific to the auditory modality, since no other sensory modality allows a similar kind of qualitative switch in communication.

During auditory communication speakers often adaptively switch from voiced to whispered vocal expression depending on the context. By “voiced” expressions we here refer to the usual or common mode of vocalizations, which to a large extend are based on vocal cord vibrations. Concerning unvoiced whispering, which does not include vocal cord vibrations, speakers not only whisper during normal speech to confine communication to nearby listeners, but they mainly switch from a voiced to a whispering mode to vocally express their inner affective states in particular contexts (Bachorowski and Owren, 2001). For example, although fearful individuals sometimes scream loudly when facing immediate danger, they most often whisper fearfully and vigilantly in threatening situations when the source of threat is not clearly detectable. Furthermore, although individuals usually express aggressive vocalizations in a voiced mode in order to intimidate another individual, they sometimes express their aggression in a low and whispered manner, especially in unfamiliar social contexts.

The switch in expression mode thus has an adaptive function, depending on certain conditions and emotional states. Alhough voiced and whispered vocalizations can express the same emotional states, they are considerably different in terms of their acoustic profile, especially in their spectral acoustic properties (Jovicic, 1998). Whispered vocalizations result from aperiodic and turbulent airflow in the vocal tract, leading to reduced salience of the vocal pitch in whispered voices (i.e.

* Corresponding author at: University of Zurich, Department of Psychology, Binzmühlestrasse 14, Box 18, 8050 Zurich, Switzerland.
E-mail address: sascha.fruehholz@uzh.ch (S. Frühholz).

http://dx.doi.org/10.1016/j.neuroimage.2016.08.023
1053-8119/

the breathiness of whispered voices) (Higashikawa et al., 1996), which usually carries important acoustic information about their affective meaning (Banse and Scherer, 1996). Besides vocal pitch, whispered and voiced vocalizations also differ in temporal features (Schwartz, 1967). This reduced acoustic profile of whispered voices, especially a reduced pitch salience, usually imposes strong challenges on the brain of human listeners. We accordingly aimed at providing a combined description, first, of the acoustic properties and, second, of the perceptual and neural network dynamics during the decoding of natural whispered and voiced affective vocalizations. These descriptions are related to our two main research questions. First, can emotions be accurately perceived in whispered voices? Second, are there similar or different neural mechanisms for decoding emotions in voiced and whispered vocalizations given that especially the latter portray only limited acoustic information?

Concerning the latter question, the neural decoding of emotions conveyed by voices predominantly, but not exclusively, involves a neural network consisting of the lateral and medial frontal cortex as well as the auditory cortex (Frühholz and Grandjean, 2013a, b; Frühholz et al., 2014), with a strong structural (Frühholz et al., 2015a) and functional connectivity (Frühholz and Grandjean, 2012) between these regions (Ethofer et al., 2012). The functional role of the auditory cortex has been proposed to underlie the acoustic analysis of physical voice features (Frühholz et al., 2012; Wiethoff et al., 2008) and the perceptual integration of voice features into an acoustic percept (Frühholz et al., 2012). Given a sufficient amount of relevant acoustic voice information, such as the level and the temporal dynamics of vocal pitch, pitch salience, intensity, or the harmonics-to-noise ratio (Frühholz et al., 2016b; von Kriegstein et al., 2010), and given the extraction of these voice features in the auditory cortex (Lewis et al., 2012; Lewis et al., 2009; Patterson et al., 2002; Penagos et al., 2004), the auditory cortex might also perform some generic emotional analysis on these voice features and the voice percept without functional support from other brain regions (Frühholz et al., 2016a). This case might be expected for the neural processing of voiced emotional vocalizations given their wide range of distinctive and discriminative acoustic features (Banse and Scherer, 1996).

However, in case of the impoverished sound quality of whispered vocalizations, the extraction of the available acoustic feature information in the auditory cortex might not be sufficient for an accurate emotional classification of emotional vocalizations, and this more challenging decoding might be supported by additional activation in an extended brain network, especially consisting of frontal brain regions. The inferior frontal cortex (IFC) shows activity and functional connectivity to the auditory cortex if acoustic cue salience in emotional voices decreases (Leitman et al., 2010), pointing to an enhanced cognitive evaluation in the IFC under challenging acoustic conditions (Frühholz and Grandjean, 2013b). The IFC might also enrich the perception of impoverished sounds by retrieving acoustic memory information from long-term, stored prototype emotional vocalizations (Binder et al., 2009). In the present study, we accordingly aimed at investigating the neural dynamics of processing voiced and whispered emotional vocalizations. We specifically investigated the acoustic decoding of these vocalizations in auditory cortical regions that are sensitive to spectral and temporal information of sounds as well as in an extended neural network in the frontal cortex that might provide in-depth cognitive evaluation to compensate for the lack of sensory sound information especially in whispered voices.

Materials and methods

Participants

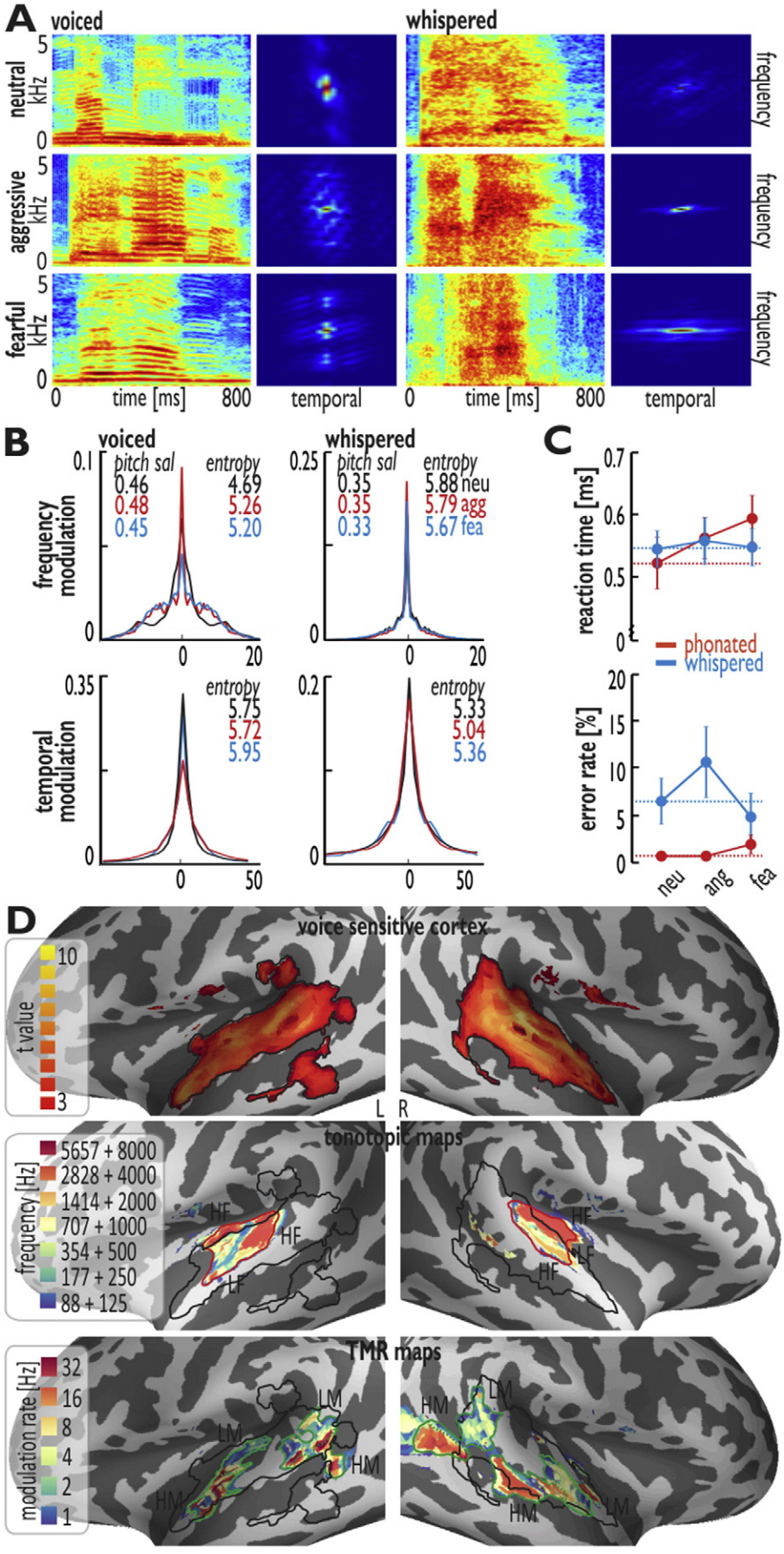
Fifteen healthy participants recruited from the Geneva University took part in the experiment (seven male; mean age 23.67 years, $SD = 3.87$, age range 18–33 years). All participants were right-handed, had normal or corrected-to-normal vision, and normal hearing abilities. No participant presented a neurologic or psychiatric history. All participants gave informed and written consent for their participation in accordance with ethical and data security guidelines of the University of Geneva. The study was approved by the local ethics committee of the University of Geneva.

Stimulus material and trial sequence

The stimulus material consisted of four speech-like but semantically meaningless two-syllable words ("*belam*", "*nolan*", "*minad*", "*namil*") spoken either in a neutral, fearful or aggressive tone (factor emotion) by two male and two female speakers including both normally voiced vocalizations and whispered vocalizations (factor phonation type), resulting in 96 different stimuli. The whispered vocalizations were produced only by exhalation from the lungs. Auditory stimuli had a mean duration of 633 ms ($SD = 172$ ms) with similar duration for neutral, aggressive and fearful voices (repeated measures ANOVA (rmANOVA); $F_{1,31} = 1.48$, *n.s.*). Stimuli were equated for mean energy ($M_{erg} = 2.98 \times 10^{-3}$ Pa^2/s, $SD_{erg} = 9.81 \times 10^{-4}$), scaled to mean sound pressure level of 70 dB, and had a linear fade-in/fade-out of 15 ms. A pre-evaluation of the stimuli by 21 participants (ten male; mean age 25.57 years, $SD = 3.58$, age range 22–34 years) revealed that neutral, aggressive and fearful voices were significantly rated as neutral (rmANOVA; $F_{4,124} = 114.86$, $p < 0.001$), aggressive ($F_{4,124} = 92.30$, $p < 0.001$), and fearful ($F_{4,124} = 192.53$, $p < 0.001$), respectively. Each stimulus had to be rated on five continuous scales (range 0–100, corresponding to low to high) on how much they expressed the respective emotion (fearful, aggressive, happy, sad, and neutral). Each stimulus was also rated for its arousal level (range 0–100, corresponding to low to high), and aggressive and fearful voices did not differ in arousal ratings (paired *t*-test; $t_{31} = 0.76$, *n.s.*), but were significantly higher in arousal ratings compared with neutral voices (aggressive: $t_{31} = 10.13$, $p < 0.001$; fearful: $t_{31} = 15.47$, $p < 0.001$).

During the main experiment, each vocalization was repeated twelve times, and they were presented in short blocks of six voices separated by 500 ms within these blocks, while the short blocks were followed by 4000 ms of no auditory stimulation. Blocks were preceded by a fixation cross of 1000 ± 175 ms duration, which cued the onset of a new block and remained on the screen until the offset of the last auditory stimulus in a block. The short blocks had a mean duration of 6359 ms (*SD* 589) and consisted of randomly chosen vocalizations of the same valence with no more than two times the same speaker or word presented in one block. After each short block participants had to indicate the valence of the vocalizations by a three alternative forced choice decision ("neutral", "aggressive", or "fearful"; response buttons counterbalanced across participants) using the index, middle, and ring finger of their right hand. The experiment was divided in two runs, and each run started either with 48 short blocks of voiced vocalizations followed by 48 blocks of whispered vocalizations, or vice versa (the order was counterbalanced across participants). Voiced and whispered vocalizations were not presented intermixed in order to avoid carry-over effects of perceptual processing from voiced vocalizations that might have influenced the perception of subsequent whispered vocalizations, or vice versa. During scanning, both voiced and whispered vocalizations were presented binaurally with magnetic resonance imaging-compatible headphones (Sensimetrics® insert earphones; http://www.sens.com/products/model-s14/) at a sound pressure level (SPL) of approximately 70 dB.

To localize human voice-sensitive regions in the bilateral superior temporal cortex (STC), we used 500 ms sound clips consisting of 70 nonhuman vocalizations and sounds (animal vocalizations, artificial sounds, natural sounds) and 70 human speech and nonspeech vocalizations presented at 70 dB SPL. The stimuli were the same as used by Capilla and colleagues (Capilla et al., 2013). Each sound clip was presented randomly once with a fixation cross preceding the onset by


A
voiced
whispered
neutral
aggressive
fearful
kHz
5
0
frequency
time [ms]
0
800
temporal
B
voiced
whispered
pitch sal
entropy
0.46
4.69
0.48
5.26
0.45
5.20
0.35
5.88 neu
0.35
5.79 agg
0.33
5.67 fea
frequency modulation
temporal modulation
5.75
5.72
5.95
5.33
5.04
5.36
C
reaction time [ms]
phonated
whispered
error rate [%]
neu
ang
fea
D
voice sensitive cortex
t value
10
3
L
R
tonotopic maps
frequency [Hz]
5657 + 8000
2828 + 4000
1414 + 2000
707 + 1000
354 + 500
177 + 250
88 + 125
HF
LF
TMR maps
modulation rate [Hz]
32
16
8
4
2
1
LM
HM

500 ms and a followed by a jittered 3550–5000 ms silent gap before the onset of the next stimulus. During the voice localizer scan, 10 randomly chosen sounds were repeated twice in a row, and participants had to detect and indicate this repetition by an index finger button press. These repetition trials served to maintain the participants' attention.

For the purpose of a tonotopic mapping of the AC, we used pure sine wave tones (PSTs) presented in an ascending progression of fourteen 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). This tonotopy localizer was identical to the one used by Da Costa et al. (2011). Pure tone bursts of a random rhythmic pattern of single PSTs (50 ms or 200 ms) of a specific frequency were presented for a random duration of 1.1–1.7 s every 2 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, 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.

Finally, for the purpose of a temporal modulation rate (TMR) mapping in auditory cortex we used amplitude modulated (AM) uniform and broadband white noise stimuli of 3000 ms duration, with modulation rates of 1, 2, 4, 8, 16, and 32 Hz, a modulation depth of 90%, and random phase of the modulation at stimulus onset. This procedure was similar to a TMR localizer used in a previous neuroimaging study (Herdener et al., 2013). Stimuli were presented in a series from low to high modulation rate separated by 1000 ms, and the series were separated by 3000 ms. These series were repeated 15 times, and stimuli were presented at 70 dB SPL.

Acoustic analysis of vocalizations

All vocal stimuli of the main experiment were first analyzed according to their level of spectral and temporal information. For this purpose, we determined the modulation spectrum of each vocalization (Fig. 1A) by converting the amplitude waveforms of the vocalizations to their spectrogram using a Hamming window of 1.81/BW window length (bandwidth (BW) 100 Hz, frequency range 1–5000 Hz), which was then filtered with a 2D-Gaussian filter (*SD* 1.3, size 3), and subjected to a 2D Fourier transform (zero-padded to match the length of the longest stimulus), from which the power was computed and the zero-frequency components were shifted to the center of the spectrum. The resulting modulation spectra for each stimulus were both averaged along the time axis and the frequency axis to give an average representation of the frequency and temporal modulation, respectively (Fig. 1B). For each stimulus we computed the entropy (i.e. Shanon entropy; Shannon, 1948) of the modulation spectrum along the temporal and the spectral axis as an indicator of the respective modulation complexity. Entropy measures for spectral information and for temporal information were separately subjected to a 2 × 3 rmANOVA with the within-subject factors *emotion* (neutral, aggressive, fearful) and *phonation type* (voiced, whispered).

Second, all vocal stimuli were analyzed according to their level of pitch salience using the Auditory Toolbox (https://engineering.purdue.edu/~malcolm/interval/1998-010/). The mean pitch salience was computed on the extracted pitch time course (derived from an autocorrelation of the original waveform) that was subjected to a cochlear model (i.e. Lyon's Passive Ear Model). The pitch salience scores were separately subjected to a 2 × 3 rmANOVA with the within-subject factors *emotion* (neutral, aggressive, fearful) and *phonation type* (voiced, whispered).

Image acquisition and analysis

We recorded imaging data on a 3-T SIEMENS Trio® TIM System (Siemens, Erlangen, Germany) by using a T2*-weighted multiplexed echo-planar imaging sequence (Feinberg et al., 2010) with an acceleration factor of four. We used a partial volume acquisition of 28 slices (thickness/gap = 2/0.4 mm, field of view [FoV] = 192 mm, in-plane resolution 2 × 2 mm, flip angle = 54°) aligned oblique to the AC-PC plane (~30° rotation) to cover all parts of the mid and inferior frontal cortex, large portions of the superior temporal cortex from anterior to posterior, and the amygdala (see Fig. S1). The sequence had a time to repetition (TR)/time to echo (TE) of 650/30 ms. This fast volume acquisition together with a short block design was chosen to obtain an improved sampling rate of the BOLD signal in order to perform a functional connectivity analysis that depends on the sampling rate of the acquired BOLD time series (see below). The functional connectivity analysis was also dependent on a continuously sampled BOLD signal. Thus, we decided to use a continuous sampling protocol instead of a sparse temporal sampling protocol (Hall et al., 1999) that are sometimes used in auditory experiments. The scanner noise did not considerably impair the recognizability of the voice stimuli, since classification accuracy was high above change level (=33%; see below). Finally, a high-resolution magnetization prepared rapid acquisition gradient echo (MPRAGE) T1-weighted sequence (192 contiguous 1 mm slices, TR/TE/time to inversion (TI) = 1900/2.27/900 ms, FoV 296 mm, in-plane resolution 1 × 1 mm) was obtained in sagittal orientation to record structural brain images from each participant.

We used the statistical parametric mapping software (SPM 8; Wellcome Trust Centre for Neuroimaging, London, UK; www.fil.ion.ucl.ac.uk/spm) for preprocessing and statistical analysis of all functional images. Functional images were realigned and 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 based on segmented grey and white matter tissue classes and used for normalizing T1 and EPI images. Functional images were spatially smoothed using an isotropic Gaussian kernel of 6mm^3 FWHM.

We used a general linear model for the first-level statistical analyses of the functional data. For the data of the main experiment, we modeled activity for each condition including boxcar functions defined by the onset and duration of the short blocks with a correct response of participants. These boxcar functions were convolved with a canonical hemodynamic response function. Separate regressors were created for each experimental condition and the general linear model also included one additional regressor containing all erroneous and missed trials. Six motion correction parameters and four regressors modeling respiratory and cardiac activity of the participants were finally included as regressors of no interest to minimize false positive activations that were due to task-correlated motion and to physiological noise. Respiratory and cardiac activity was modeled using the approach of Glover and colleagues (Glover et al., 2000) referred to as the RetroIcor algorithm (code provided by http://cbi.nyu.edu/software/).

Fig. 1. Voice stimuli and functional localizer scans. (A) Spectrograms and modulation spectra for voiced and whispered neutral (neu), aggressive (agg), and fearful vocalizations (fea). (B) Mean modulation spectrum along the frequency and the temporal dimension of the 16 stimuli for each condition, including measures of mean pitch salience ("pitch sal") and entropy. (C) Reaction times and percentage error rates across the participants ($n = 15$); mean ±1SEM. (D) Extended AC/STC activity for vocal relative to nonvocal sounds (top panel; *t* contrast); tonotopic maps with a common high-low-high gradient (high frequency, HF; low frequency, LF; *F* contrast) (middle panel); and anterior and posterior temporal modulation rate (TMR) maps with a superior (low modulation rate, LM) to inferior orientation (high modulation rate, HM) (lower panel; *F* contrast). All activations across the participants ($n = 15$) are thresholded at $p = 0.001$ with a cluster extent of $k = 33$ voxels, corresponding to $p < 0.05$ corrected at the cluster level.

Linear contrasts for the experimental conditions for each participant were taken to a second-level random effects analysis implemented as a factorial design including the within-subject factors emotion (neutral, anger, fear) and phonation type (voiced, whispered). Functional activations resulting from the comparison between conditions were thresholded at a voxel-level threshold of $p = 0.001$ and a cluster extend threshold of $k = 33$. This combined voxel and cluster threshold corresponds to $p = 0.05$ corrected at the cluster level and was determined by the 3DClustSim algorithm implemented in the AFNI software (http://afni.nimh.nih.gov/afni) according to the estimated smoothness of the data across all contrasts. The cluster extent threshold of $k = 33$ was the maximum value for the minimum cluster size across contrasts of the main experiment and of functional localizer scans.

In order to determine a directed functional connectivity analysis between regions, which showed activation either for the comparison of neutral and emotional vocalizations (factor *emotion*) or for the comparison of voiced with whispered vocalizations (factor *phonation type*), we performed a granger causality analysis (GCA). Compared to other methods of estimating functional connectivity, such as the psychophysiological interaction (PPI) analysis (Friston et al., 1997), GCA has the advantage of providing information about the directionality of functional connections. Furthermore, directionally of connectivity estimation in GCA more accurately and validly can be applied to blocked presentations of stimuli, while directionality estimation in dynamic causal modeling (DCM) seems more effective for event-related designs (Seth et al., 2015). Thus, for the GCA analysis concerning the factor *emotion* we computed both the functional connectivity for all regions that were more sensitive to neutral relative to emotional vocalizations and for all regions that were more sensitive to emotional relative to neutral vocalizations (Fig. 2A). The latter analysis also included all regions for which we found an interaction between the factors *emotion* and *phonation type* (Fig. 2D). For the GCA analysis concerning the factor *phonation type* we computed functional connectivity for all the regions that were more sensitive to voiced relative to whispered vocalizations and for all regions that were more sensitive to whispered relative to voiced vocalizations (Fig. 2B). Both connectivity analysis for the factor *phonation type* also included all regions, which showed either an interaction including voiced or whispered vocalizations, respectively (Fig. 2D).

The GCA analysis was performed on time intervals spanning from one volume acquisition before the onset (i.e. to take into account some pre-stimulus neural activity given that a short block started randomly during the following volume acquisition) of a short block until 25 volumes after onset. The amount of 25 volumes after the short block onset was chosen taking into account the mean short block length (6.359 s) as well as the average peak BOLD delay (~5 s) and its relaxation to baseline (~5 s), resulting in post-stimulus interval of ~16.359 s. These segments were extracted from the entire time course of activity in each region extracted as the first eigenvariate in a 2 mm radius sphere around voxel of peak activations. Extracted segments were pooled across subjects. We used the Granger Causal Connectivity Analysis toolbox (Seth, 2010) to perform this GCA analysis. The GCA analysis is based on vector autoregressive (VAR) models, and the directional connections strengths between brain regions is quantified as the log ratio of predictions errors. A higher log ratio level for a certain connection is indicative of a greater causal influence (Seth et al., 2015). Segmented intervals were preprocessed by linear detrending and by removing the temporal and the ensemble mean. All data showed covariance stationarity as indicated by the KPSS test, which is a test with the null hypothesis of no unit root. Finally, a model order of two was chosen based on previous studies with approximately half the sampling rate using a model order of one (Deshpande et al., 2009; Sato et al., 2010), and model validity was confirmed by the Durbin-Watson test. Significant directed connections based on *F* statistics between brain regions were determined at a threshold of $p = 0.0001$ (FDR corrected).

On the resulting GCA data we subsequently performed a Graph Theoretic Analysis (GTA) using the Brain Connectivity Toolbox (Rubinov and Sporns, 2010). For each region, we first determined the node strength in terms of the sum of weights of all incoming, all outgoing, and of all connections in total for each node (i.e. region). Second, we determined the connection weight for all connections, which survived the thresholding in the GCA analysis. For the illustration of the connectivity results of the GCA analysis, we used the CIRCOS software (http://circos.ca/software/). Additionally, we computed lobe specific connectivity measures, by determining the mean weights of connection in the same lobe (self), connections to the same lobe in the contralateral hemisphere (horizontal), connections to the other ipsilateral lobe (vertical), and connections to the other contralateral lobe (diagonal).

For the voice localizer, single-subject activity was modeled for vocal and nonvocal stimuli as stick functions defined by the onset of the auditory stimuli, convolved with a canonical hemodynamic response function. Repetition trials were modeled as a separate regressor of no interest. Linear contrasts for the vocal and the nonvocal conditions for each participant were taken to a second-level random effects analysis. For the latter, we contrasted all vocal against nonvocal stimuli at a threshold of $p = 0.001$ and a cluster extent of $k = 33$ voxels (see above). We determined voice-sensitive regions along the superior temporal gyrus (STG) and superior temporal sulcus (STS) in both hemispheres.

For the tonotopic localizer scan, we modeled activity for each condition as boxcar functions defined by the onset and duration of each PST level, convolved with a canonical hemodynamic response function. Linear contrasts for each condition for each participant were taken to a second-level random effects analysis, which was setup as 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. By using an *F* contrast across these different conditions we determined areas in the auditory cortex that showed a significant difference between these conditions at a combined threshold of threshold of $p = 0.001$ and a cluster extent of $k = 33$ voxels (see above). 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.

For the TMR localizer scan, we modeled activity for each condition as boxcar functions defined by the onset and duration of each AM level, convolved with a canonical hemodynamic response function. Linear contrasts for each condition for each participant were taken to a second-level random effects analysis, which was setup as a 1×6 factorial design, using the thresholding strategy from the tonotopic localizer.

Results

Acoustic analysis of vocalizations

The present study included a total of 96 affective voices expressing neutral, aggressive, and fearful emotions in either a voiced or a whispered mode (Fig. 1A). We first measured their pitch salience and the modulation spectrum as the most distinguishing acoustic features. The modulation spectrum assesses spectral and temporal regularity, and the entropy of this spectrum quantifies the amount of spectral and temporal information (Fig. 1B, Table S1).

The entropy for spectral information differed across the *phonation types* (repeated measures ANOVA (rmANOVA); $F_{1,15} = 75.21$, $p < 0.001$; $M_{phon} = 5.04 < M_{whis} = 5.78$), but not across *emotions* ($F_{2,30} = 3.43$, $p = 0.045$; but Bonferroni posthoc planned pairwise comparison were nonsignificant, all p's > 0.078; $M_{neu} = 5.29$, $M_{ang} = 5.52$, $M_{fea} = 5.44$). There was an *emotion* by *phonation type* interaction ($F_{2,30} = 8.07$, $p = 0.0015$). For the latter, follow-up two-sided paired *t*-test indicated that for voiced vocalizations aggressive ($t_{15} = 4.87$, $p < 0.001$) and fearful vocalizations ($t_{15} = 3.31$, $p = 0.0046$) had higher spectral entropy relative to neutral vocalizations, but no such

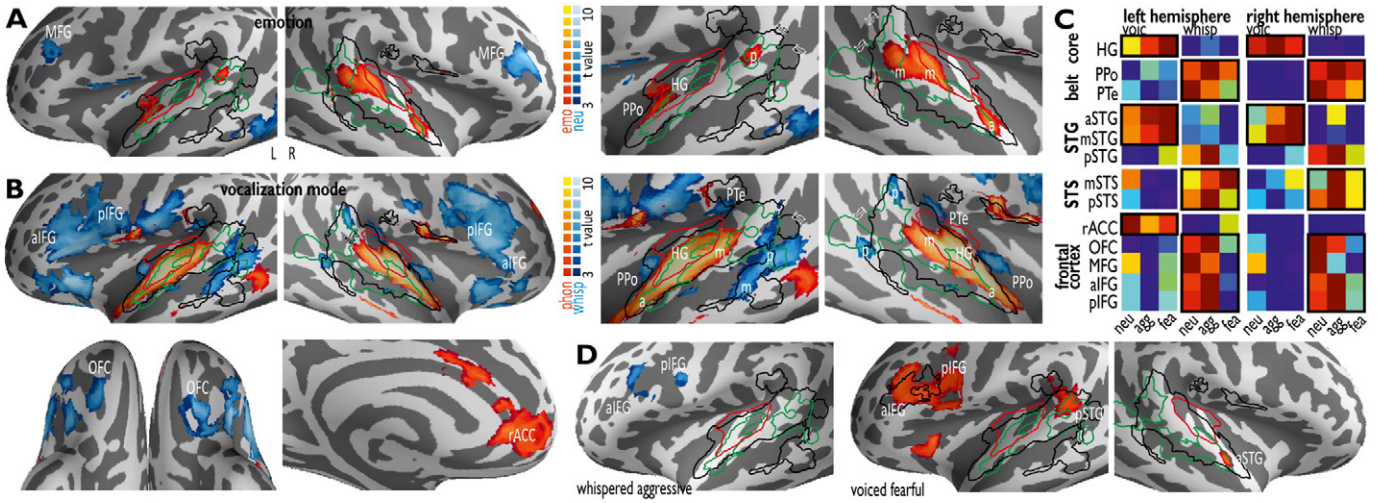


Fig. 2. Functional activity for voiced and whispered affective voices. (A) Activity for affective relative to neutral vocalizations (red) and for neutral relative to affective vocalizations (blue) across participants ($n = 15$). The right panel shows an enlarged view of activity in the temporal cortex. (B) Functional activity for voiced relative to whispered vocalizations (red) and for whispered relative to voiced vocalizations (blue). (C) Beta estimates for several left and right hemispheric subregions extracted as the mean signal in a 2 mm radius sphere around voxel of peak activations. The beta estimates are scaled to the maximum in each subregion, that is, the core, belt, STG, STS, frontal cortex region. (D) Interaction effects for whispered aggressive vocalizations (left panel) in the left IFG and for voiced fearful vocalizations (right panel) in the left IFG, OFC, pSTG, and right aSTG. All activations across the participants ($n = 15$) are based on t contrasts and thresholded at $p = 0.001$ and $k = 33$, corresponding to $p < 0.05$ corrected at the cluster level. Abbreviations: *Ins* insula, *MFG* middle frontal gyrus, *voic* voiced, *STS* superior temporal sulcus, *whisp* whispered.

differences were observed with whispered vocalizations (all t_{15}'s < 2.03, all p's > 0.06). The entropy for temporal information differed across the different *phonation types* (rmANOVA; $F_{1,15} = 63.26$, $p < 0.001$; $M_{phon} = 5.81 > M_{whis} = 5.25$) and across *emotions* ($F_{2,30} = 8.16$, $p = 0.0014$; $M_{ang} = 5.38 < M_{neu} = 5.54$ and $M_{fea} = 5.66$), but there was no *emotion* by *phonation type* interaction ($F_{2,30} = 1.59$, *n.s.*).

The pitch salience of vocal stimuli differed across the *phonation types* (rmANOVA; $F_{1,15} = 146.83$, $p \leq 0.001$; $M_{phon} = 0.46 > M_{whis} = 0.35$), but not across *emotions* ($F_{2,30} = 1.41$, *n.s.*). There was no *emotion* by *phonation type* interaction ($F_{2,30} = 0.26$, *n.s.*).

Behavioral data

To investigate the neural dynamics during the processing of these vocalizations, we asked 15 healthy human participants to listen to and classify these vocalizations (as neutral, aggressive, or fearful) during a high-spatial and high-temporal resolution functional magnetic resonance imaging study. Reaction times of these classifications did not differ across *emotions* (rmANOVA; $F_{2,28} = 1.97$, *n.s.*) and *phonation types* ($F_{1,14} = 0.17$, *n.s.*), and there was no *emotion* by *phonation type* interaction ($F_{2,28} = 2.24$, *n.s.*) (Fig. 1C). Error rates did not differ across *emotions* (rmANOVA; $F_{2,28} = 0.77$, *n.s.*), but across the different *phonation types* ($F_{1,14} = 15.80$, $p = 0.0013$). Participants made more errors in classifying whispered relative to voiced vocalizations, but their accuracy was still at a high level (>90%) (Fig. 1C). There was no *emotion* by *phonation type* interaction ($F_{2,28} = 1.44$, *n.s.*).

Functional localizer scans

The experiment included three different localizer scans to identify cortical auditory regions that show tonotopic frequency mapping (i.e. sensitivity to spectral information), that are sensitive to the temporal modulation rate (TMR) of sounds (i.e. sensitivity to temporal information), or that show sensitivity to human voices. In terms of neural activations for the localizer and the experimental scans, we determined functional effects in subregions of the auditory cortex (AC) consisting of auditory "core" (Heschl's gyrus, HG) and "belt" regions (planum polare (PPo), planum temporale (PTe)) (Hackett, 2011). We also determined activity in subregions of the STC consisting of STG and STS (Frühholz and Grandjean, 2013a). During the main experiment (see below) we also determined activity in frontal lobe subregions especially of the inferior frontal gyrus (IFG) and orbitofrontal cortex (OFC) that are central for the decoding of vocalizations (Frühholz and Grandjean, 2013b).

During the voice localizer scan, we presented human vocal and non-human sounds. Vocal compared to nonvocal sounds produced extended activity in bilateral STC covering regions of core and belt auditory regions as well as of anterior-to-posterior STG and STS (Fig. 1D, upper panel). Regions showing voice sensitivity are marked by a black outline in Fig. 1–2.

To identify tonotopic cortical fields, a tonotopy localizer scan included the auditory stimulation with pure tones between 0.088 and 8 kHz and revealed a typical pattern of frequency maps with a mirror-symmetric frequency progression (high-low-high) (Da Costa et al., 2011), with the low-frequency area approximately located along left Heschl's sulcus and right PTe (Fig. 1D, middle panel). These two maps are thought to correspond to regions *A*1 and R (rostral area) of the primate primary auditory core. Regions showing tonotopic mapping are marked by a red outline in Fig. 1–2.

Finally, to identify cortical regions sensitive to the temporal amplitude modulation rate (TMR) of sounds, a TMR localizer scan was used that included the auditory stimulation of white noise stimuli with intensity variations in the range of 1–32 Hz (see Giraud et al., 2000; Herdener et al., 2013). This scan revealed two bilateral amplitude modulated (AM) gradient maps located in posterior and mid STC. The posterior map showed a low-to-high AM frequency progression in the direction of STG-STS to middle temporal gyrus MTG. The left anterior map showed a low-to-high AM frequency progression in the direction of mHG-STG (i.e. mid HG to STG), while the right anterior map showed a slightly more lateral lHG-mSTS (i.e. lateral HG to mid STS) progression. Regions showing AM mapping are marked by a green outline in Fig. 1–2.

Functional brain activity of the main experiment

In terms of the affective value of the vocalizations, we first compared neural activity for affective relative to neutral vocalizations (i.e. the emotion effect). They elicited activity in the left core and belt regions and in bilateral higher-level auditory regions in the STC (Fig. 2A, Table 1). The activity was specifically found in left HG and PPo as well as in right anterior STG (aSTG), mid STG (mSTG), and mid STS (mSTS).

Table 1
Functional peak activity related to the affective value of the vocalizations. (A) Activity for affective versus neutral vocalizations. (B) Activity for neutral versus affective vocalizations. Activations are reported in the left ("L") and the right hemisphere ("R"). Activations are thresholded across all participants ($n = 15$) at $p = 0.001$ with a cluster extent of $k = 33$ voxels, corresponding to $p < 0.05$ corrected at the cluster level. Peak coordinates with no cluster size represent local peaks within the larger cluster that is defined by the cluster size mentioned above it.

Region	Cluster size	z value	MNI		
			x	y	z
(A) Affective vs. neutral vocalizations					
L Heschl's gyrus	79	4.67	−46	−12	0
L planum polare		3.69	−40	−20	0
L superior temporal gyrus	42	4.11	−64	−42	16
R superior temporal gyrus	445	5.29	66	−18	8
		4.90	66	−24	12
		4.77	60	−38	8
R superior temporal sulcus		4.66	58	−26	10
		4.26	68	−24	6
R superior temporal gyrus	125	4.91	62	2	−8
		4.44	60	0	0
R superior temporal gyrus		3.41	50	0	−4
R Heschl's gyrus		3.71	48	−8	2
(B) Neutral vs. affective vocalizations					
L middle frontal gyrus	39	3.61	−42	32	26
L inferior temporal gyrus	42	3.48	−44	−56	−18
		3.39	−34	−52	−16
		3.31	−44	−66	−16
L frontal operculum	62	3.64	−40	0	14
R middle frontal gyrus	346	5.35	44	42	20
		4.91	46	46	16
L middle temporal gyrus	71	3.88	−50	−56	−4
L inferior occipital gyrus	40	4.06	−16	−96	−10
L middle occipital gyrus	135	3.75	−24	−90	0
R inferior occipital gyrus	51	3.86	18	−92	−6

Neutral relative to affective voices elicited activity in bilateral middle frontal gyrus (MFG).

In terms of expression mode, voiced relative to whispered vocalizations revealed auditory cortical activity in the voice-sensitive cortex in the bilateral core regions (HG) and mid-anterior STG, which largely overlapped with the tonotopic maps (Fig. 2B, Table 2). Whispered relative to voiced vocalizations elicited activity in the bilateral voice-sensitive cortex, especially in the belt regions (PPo) of posterior STC. Besides these differential auditory cortex activations for voiced vocalizations and whispered vocalizations, both modes of vocalization also revealed activity in the frontal cortex, which was confirmed by signal extracted from several regions of interest (Fig. 2C). Voiced vocalizations elicited activity in the rostral anterior cingulate cortex (rACC), whereas whispered vocalizations elicited activity in bilateral OFC and the lateral inferior frontal cortex (IFC) consisting of anterior IFG (aIFG) and posterior IFG (pIFG).

An interaction analysis also revealed specific frontal activity, such as activity of the left aIFG and pIFG for whispered aggressive vocalizations and voiced fearful vocalizations, with additional activity for the latter in bilateral superior temporal gyrus (left pSTG, right aSTG) (Fig. 2D, Table 3). No functional brain activity based on an interaction analysis was found for whispered fearful vocalizations and for voiced aggressive vocalizations.

Directional functional connectivity

To identify the dynamics in the neural network described above, we performed a directional functional granger causality analysis separately for the regions decoding neutral or affective vocalizations (Fig. 3A–B) and for those decoding voiced or whispered vocalizations (Fig. 3C–D). We first compared the neural network for neutral and affective vocalizations. Unlike the sparse functional network for neutral vocalization connecting right to left MFG (Fig. 3A, left panel), affective vocalizations

Table 2
Functional peak activity related to the vocal expression mode. (A) Activity for voiced versus whispered vocalizations. (B) Activity for whispered versus voiced vocalizations. Activations are thresholded across all participants ($n = 15$) at $p = 0.001$ with a cluster extent of $k = 33$ voxels, corresponding to $p < 0.05$ corrected at the cluster level. *Inf* infinite.

Region	Cluster size	z value	MNI		
			x	y	z
(A) Voiced vs. whispered vocalizations					
L Heschl's gyrus	1225	Inf	−52	−14	6
L superior temporal gyrus		Inf	−60	−10	4
		7.47	−62	−26	14
		6.70	−58	2	−4
L middle temporal gyrus	101	4.62	−60	−62	−2
R Heschl's gyrus	1246	Inf	54	−12	4
R superior temporal gyrus		Inf	66	−18	8
		5.86	62	3	−8
L cingulate gyrus	372	4.59	−6	36	2
		4.42	−4	46	4
L cingulate gyrus	141	4.34	−4	26	26
R cingulate gyrus	106	4.04	2	−10	32
R superior frontal gyrus	100	4.57	20	56	32
R middle frontal gyrus	44	3.82	14	−58	12
L caudate nucleus	256	5.26	−18	−4	26
	47	3.97	−12	20	16
R caudate nucleus	41	4.69	20	−10	26
	46	4.29	18	−10	14
L parahippocampal gyrus	163	4.84	−24	−32	−18
L hippocampus		4.73	−26	−30	−6
L fusiform gyrus	41	3.81	−24	−48	−22
L thalamus	134	4.14	−2	−12	12
L thalamus (MGN)		3.82	−14	−28	−4
R thalamus (MGN)	41	3.78	14	−28	2
R thalamus		3.55	18	−26	−6
(B) Whispered vs. voiced vocalizations					
L planum polare	58	5.08	−40	−22	−4
L superior temporal sulcus	157	4.83	−30	−36	10
L planum temporale		3.99	−40	−38	24
		3.98	−34	−40	16
L superior temporal sulcus	197	4.17	−56	−54	6
		4.17	−56	−38	−4
		3.35	−54	−62	20
R planum temporale	91	4.95	42	−36	16
R superior temporal gyrus	111	4.86	66	−42	22
		4.12	64	−50	20
R superior temporal sulcus		3.91	64	−52	12
R planum polare	47	4.44	42	−20	−4
L inferior frontal gyrus	1579	6.36	−32	28	0
		4.94	−52	6	22
		4.22	−42	22	24
L orbital gyrus	98	5.38	−20	26	−20
R inferior frontal gyrus	1438	6.01	38	14	30
		5.58	52	22	16
		4.71	48	26	2
R inferior frontal gyrus	37	4.62	44	38	−12
R orbital gyrus	53	4.07	20	30	−14
R precentral gyrus	40	3.98	66	−6	28
L postcentral gyrus	213	5.62	−60	−16	18
L inferior occipital gyrus	78	4.76	−20	−68	−28
L cuneus	50	3.96	−12	−84	−12
L inferior occipital gyrus		3.59	−18	−78	−6
R inferior occipital gyrus	73	4.94	22	−78	−12
R hippocampus	46	4.72	26	−38	18

revealed a widespread temporo-frontal network connecting left IFC, left auditory core/belt regions, and bilateral STG (Fig. 3A, right panel). This affective vocalization network indicated a relatively strong importance of the right temporal cortex (Fig. 3B, lower left panel), and especially of the aSTG both as input and output node (Fig. 3A, right panel). Furthermore, concerning the important role of the right temporal cortex in affective vocalization network, the connections weights of the right temporal cortex were higher for whispered (Fig. 3B, lower right panel) compared to voiced affective vocalizations (Fig. 3B, lower mid panel). The central role of the right temporal cortex in the affective vocalization network was also indicated by lobe-specific connectivity

Table 3
Functional peak activity for the comparison of the interaction effects. (A) Specific activity for whispered angry vocalizations. (B) Specific activity for voiced fearful vocalizations. Activations are reported in the left ("L") and the right hemisphere ("R"). Activations are thresholded across all participants ($n = 15$) at $p = 0.001$ with a cluster extent of $k = 33$ voxels, corresponding to $p < 0.05$ corrected at the cluster level.

Region	Cluster size	z value	MNI		
			x	y	z
(A) Whispered × angry vocalizations					
L inferior frontal gyrus	37	3.71	−48	10	28
L inferior frontal sulcus	35	3.36	−42	26	30
		3.35	−44	22	30
(B) Voiced × fearful vocalizations					
L superior temporal gyrus	124	3.95	−62	−42	18
		3.42	−52	−42	22
		3.36	−54	−44	18
R superior temporal gyrus	35	4.25	62	0	−6
L inferior frontal gyrus	623	4.57	−46	22	28
		4.32	−50	12	28
L middle frontal gyrus		4.26	−44	30	24
L insula	93	4.38	−28	20	6
		3.89	−30	24	−2

measures as determined by the mean weights of intra- and extra-regional connections. The right temporal cortex, first, showed high connections weights for intraregional connections (i.e. "self" connections), second, for bidirectional connections with the left temporal cortex (i.e. horizontal connections), and, third, for afferent connections from the left IFC (i.e. diagonal connections). These three main network features were characteristic of the general affective vocalization network (Fig. 3B, lower left panel), and again especially for the whispered affective vocalizations (Fig. 3B, lower right panel).

A relative increase of connection weights involving the right temporal cortex was also found for the voiced vocalization network along with the left temporal cortex, with bilateral mSTG as a central input/output node (Fig. 3C). Both the left and right temporal cortex showed intraregional connections and relatively increased STC interconnections. A left frontal intra-regional and a left frontal to contralateral right temporal connection with low connection weights were also found (Fig. 3D). While the latter left frontal to right temporal connection was specific to all voiced vocalizations (Fig. 3D, top left panel), a left frontal to ipsi-lateral left temporal connection was found for aggressive and fearful voiced vocalizations (Fig. 3D, top mid panels). The later resulted from an independent estimation of functional connections of each emotion separately for voiced vocalization. Compared to the relatively simple voiced vocalization network, we found a multi-directional complex network during the recognition of the affective value from whispered vocalizations, and this whispered vocalizations network showed several unique features. First, this network generally included highly increased intra- and inter-regional connection weights of both temporal and frontal cortices (Fig. 3D, lower panel) compared to the voiced vocalization network (Fig. 3D, upper panel). Second, unlike the voiced vocalizations network, the whispered vocalization network included bottom-up auditory-to-frontal connections (i.e. vertical and diagonal connection from temporal areas), indicating the transfer

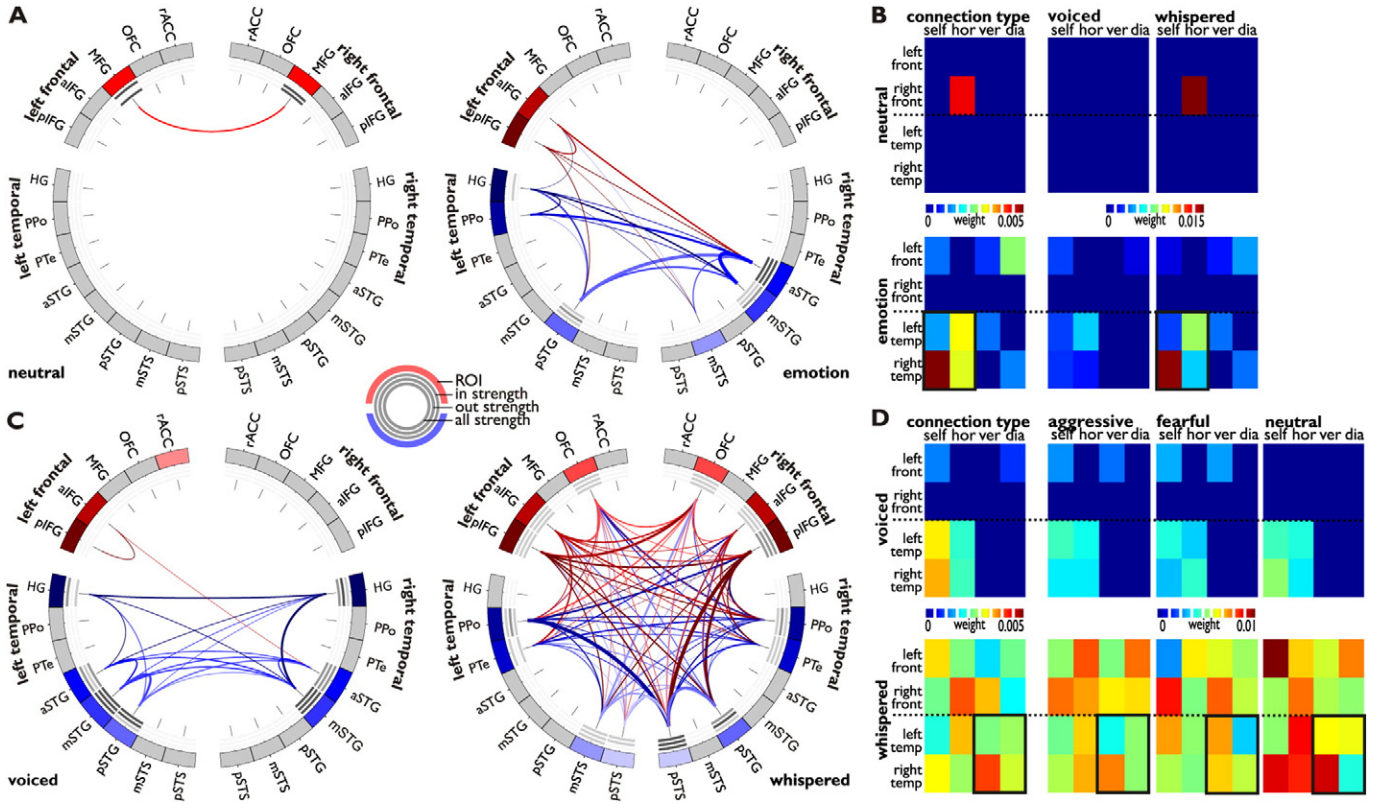


Fig. 3. Results of the granger causality and graph theoretic analysis. (A) Functional connections for neutral (left; $n = 881$ trials) and for affective vocalizations (right; $n = 1867$ trials). Seed regions: *red* frontal, *blue* temporal; connection color according to the seed region of interest; line thickness indicates connection strength; inner circles represent connections weights: incoming (in strength), outgoing (out strength), and all in/out connections (all strengths). (B) Mean connection strength for neutral and affective vocalizations (left panel) for connections inside the same lobe (*self*), to the same contralateral lobe (*hor*, horizontal), to the other ipsilateral lobe (*ver*, vertical), and to the other contralateral lobe (*dia*, diagonal). Connection strength shown separately for voiced (middle panel) and whispered vocalizations (right panel). The black boxes in the lower panel indicate the increased temporo-temporal connections in the emotional vocalizations network, which were especially increased for the whispered emotions. (C) Functional connections for voiced (left; $n = 1383$ trials) and for whispered vocalizations (right, $n = 1365$ trials). (D) Mean connection strength for voiced and whispered vocalizations (left panel) and separately for aggressive (middle left panel), fearful (middle right panel), and neutral vocalizations (right panel). The black boxes in the lower panel indicate the bottom-up temporal-to-frontal connections in the whispered vocalizations network, that were not found in the voiced vocalization network. Connections were thresholded at $p = 0.0001$ (false discovery rate (FDR) corrected).

of auditory information to the frontal cortex. Third, the whispered vocalization network also included stronger frontal-to-frontal connections (i.e. "self" and "horizontal" connection from frontal areas) as well as top-down frontal-to-auditory connections (i.e. vertical and diagonal connection from frontal areas), indicating an increased involvement of cognitive evaluation processes.

Discussion

The present study determined the neural activity and network dynamics underlying the decoding of the socio-affective value from voiced and whispered vocalizations. These vocal expression modes differ considerably in their acoustic properties, especially in terms of tonal information. Concerning the latter, voiced vocalizations had a higher pitch salience, probably based on their higher spectral regularity (i.e. less entropy) compared with whispered vocalizations. The affective classification of whispered vocalizations was slightly more difficult, probably because of this missing spectral information (Vestergaard and Patterson, 2009). This is in accordance with the observation that individuals are less confident in their decisions on whispered relative to voiced vocalizations (Lass, 1976). Unlike their reduced pitch salience, whispered vocalizations showed more temporal regularity than did voiced vocalizations. Given the impoverished tonal sound quality of whispered voices, from which no valid affective value can be inferred, increased temporal regularity of whispered vocalizations might partly compensate for this missing tonal information. Speakers might actually vocalize more regularly and with increased modulation depth to improve the clarity of whispered vocalizations (Krause and Braida, 2004), which might also improve their recognizability and their affective classification.

In terms of neural activity underlying these classifications, we first determined activity related to the affective value of vocalizations. Affective vocalizations elicited activity in the left core and belt regions, and in bilateral higher-level auditory regions in the STC. These auditory subregions have been identified as being relevant for affective voice processing (Beaucousin et al., 2007; Frühholz and Grandjean, 2013a; Frühholz et al., 2015b; Frühholz et al., 2016a). They were located in the bilateral voice-sensitive cortex, and these regions partly overlapped with low frequency fields of the tonotopic frequency maps (<3–4 kHz), and with rather low temporal regularity fields (low modulation (LM)) of the TMR maps (<4–8 Hz). This distributed activity is in accordance with previous observations (Frühholz and Grandjean, 2013a; Pernet et al., 2015), and the overlap with specific tonotopic and TMR fields reflects the importance of these frequency and temporal information ranges for classifying affective vocalizations (Banse and Scherer, 1996).

Concerning the neural activity related to the expression mode, voiced vocalizations showed activity in voice-sensitive auditory core regions and the STG, which largely overlapped with the tonotopic maps. HG activity is in accordance with the activity found for voiced relative to *synthesized* whispered vocalizations in the anterolateral HG (von Kriegstein et al., 2010), but here we report a much more extended activity in response to *natural* voiced and whispered vocalizations. The activity inside the tonotopic fields might represent the higher spectral regularity of voiced vocalizations (Norman-Haignere et al., 2013) that determines the perception of pitch. Pitch and pitch salience are predominantly used in listeners to classify affective vocalizations (Banse and Scherer, 1996; Vestergaard and Patterson, 2009). The activity for voiced vocalizations largely overlapped with the activity for the emotion effect, pointing to the relevance of salient pitch information for decoding the affective value of vocalizations (Banse and Scherer, 1996; Frühholz et al., 2012; Leitman et al., 2010). The activations for voiced vocalizations also overlapped with the left (LM and high modulation (HM) field) and the right anterior TMR map (LM field), which seems to decode both slow and fast temporal dynamic information from voiced vocalizations. This information might be related to segmental and suprasegmental modulation of intonation contours of vocalizations (Banse and Scherer, 1996).

Whispered relative to voiced vocalizations elicited activity in the bilateral voice-sensitive cortex. The effects for whispered vocalizations overlapped with the emotion effect only in the left belt region (i.e. PPo), indicating some direct affective decoding in the secondary AC from whispered vocalizations. STC activity in response to whispered vocalizations was located outside the tonotopic maps, but overlapped with the posterior TMR maps. In accordance with their reduced spectral regularity, whispered vocalizations elicited activity only in auditory regions sensitive to temporal regularity, a feature that improves their recognition accuracy (Krause and Braida, 2004). Thus, auditory cortical activity in response to whispered vocalizations might indicate both the decoding of the affective meaning (i.e. PPo) and the decoding of temporal information (i.e. posterior STS). Concerning this temporal information, temporal regularity was interestingly decoded in the anterior TMR maps for voiced vocalizations, whereas for whispered vocalizations, it was decoded in the posterior TMR maps. Given that whispered vocalizations had higher temporal regularity, this might indicate that the anterior and posterior TMR maps are differentially sensitive to temporal information in more regular and irregular sounds, respectively. This would point to a diversity of cortical TMR maps beyond their existing topographical description (Herdener et al., 2013).

Besides STC activity, both modes of vocalization also revealed activity in the frontal cortex. In response to voiced vocalizations, activity was found in rACC, which is usually involved in inferring the emotional state of another individual (Frühholz et al., 2009; Frühholz et al., 2016a; Szameitat et al., 2010). Whispered vocalizations elicited activity in the OFC and the IFC, which are involved in the cognitive evaluation of vocalizations (Frühholz and Grandjean, 2013b), especially under demanding task conditions (Leitman et al., 2010). The IFC might be also involved in retrieving information from long term semantic memory (Binder et al., 2009), such that the impoverished sound quality of whispered vocalization is enriched with acoustic memory information from previous encounters with whispered voices or from prototypical information from voiced emotional vocalizations. In the present study, participants performed less accurately when classifying whispered vocalizations. This indicated increased difficulty to recognize the affective value from whispered voices, which might lead to increased evaluation demands for correct classification in the OFC and IFC. The activity in the IFC in response to whispered vocalizations also extended into the inferior motor cortex, which usually controls articulatory vocal motor parameters. Motor cortex activity has been shown to support speech perception (Hervais-Adelman et al., 2012; Pulvermüller et al., 2006), and this motor cortex activity found here might also support the recognition of affective intonations in speech during the more challenging task of recognizing emotions from whispered vocalizations.

This role of the IFC in supporting the classification of vocalizations under more challenging conditions is also indicated by the results of an interaction analysis for whispered aggressive vocalizations and voiced fearful vocalizations. Both emotion-by-mode combinations are "unusual" ways of expressing these different emotions. Aggressive utterances are usually expressed with a prominent voice, whereas fear is often expressed in a soft and whispered tone. These unusual affective vocalizations might receive enhanced evaluation in the IFC for correct classification (Leitman et al., 2010), with voiced fear also showing increased activity in the bilateral STC and overlapping with the LM area of the TMR maps. Thus, increased decoding of slow temporal information in voiced fearful vocalizations might support their correct classification. These increased decoding demands for these rather unusual ways of emotional vocalization might also be supported by the retrieval of acoustic memory information as discussed above, which might be supported by IFC activity both for the unusual whispered and the voiced vocalizations.

The functional activity underlying these classifications thus involved a widespread neural network for the decoding of the expressions mode

and the affective value of vocalizations. Concerning these neural network dynamics as determined by a Granger causality analysis, affective vocalizations revealed a widespread temporo-frontal network with strong connectivity to right temporal auditory cortex, especially for whispered affective vocalizations. The right STC showed strong intraregional connections, but also bidirectional connectivity with the left auditory STC and top-down connections from the left frontal cortex. The right STC seems to be important in integrating and distributing information between brain regions, especially under conditions of increased decoding demands (Leitman et al., 2010), such as for whispered affective vocalizations. This finding might also explain increased task-relevant top-down modulation of the right auditory cortex by the left frontal cortex, which might drive enhanced acoustic analysis in the AC (Frühholz et al., 2016b; Leitman et al., 2010).

Concerning neural network dynamics related to the expression mode, a relative increase in connectivity of the right and the left auditory cortex was also found both for the voiced and whispered vocalizations networks. Concerning the voiced vocalizations network, bilateral core auditory cortices were interconnected pointing to increased acoustic decoding and inter-hemispheric acoustic information exchange (Frühholz and Grandjean, 2012). The bilateral mid STG, in particular, had a central position as an input and output node in this network. The mid STG only marginally overlapped with the tonotopic and the TMR maps and could thus represent a perceptual integration node of spectral and temporal acoustic information (Frühholz et al., 2012). Given only minor top-down connections from the frontal cortex for voiced vocalizations, their decoding might largely rest on an in-depth acoustic analysis in the auditory cortex. Given that voiced vocal expressions are rich in terms of spectral and time information, this acoustic decoding seems sufficient for a proper value attribution to these vocalizations (Banse and Scherer, 1996).

Unlike the voiced vocalizations network, we found a multidirectional "hypercomplex" network (Norris et al., 2005) with several unique features for the decoding of whispered vocalizations. First, we identified a much stronger involvement of the bilateral frontal regions, which showed many fronto-frontal and especially fronto-temporal top-down connections. The top-down influence of frontal regions on the auditory cortex might drive an increased acoustic analysis of the available sound information in whispered voices (Leitman et al., 2010). Second, temporal regions not only showed intra- and interregional connections, but also extended bottom-up temporo-frontal connections. These bottom-up and the top-down connections might be required to decode the affective value from whispered voices due to their impoverished acoustic quality. The auditory cortex might send the analysis of the sparsely available acoustic information to the frontal cortex, which consequently might provide increased cognitive evaluation (Frühholz and Grandjean, 2013b; Leitman et al., 2010) and enrichment by voice memory information to support top-down predictions on acoustic decoding on the auditory cortex (Bar, 2009). This temporal bottom-up transfer of acoustic information and the frontal top-down influence might be an iterative process until an accurate decision can be made on whispered vocalizations. Third, all of these temporal and frontal regions had an equal level of importance as input and output nodes in this network, except for the relatively increased importance of regions in the right posterior STC. The latter regions, located in the posterior TMR map, may decode temporal information in more regular vocalizations. Whispered vocalizations were temporally more regular and had only marginal spectral information. Thus, the functional connectivity with brain regions decoding this information seems to be of high relevance (Trost et al., 2015; Trost et al., 2014). These neural network data together suggest that a large-scale multidirectional bottom-up and top-down brain network compensates for the impoverished sound quality of whispered voices to support their accurate recognition and value attribution.

Taken together, our data demonstrate, first, that the affective value can be quite accurately recognized both from whispered and from voiced vocalizations, although to a slightly lesser degree in vocal whispering. Second, the affective value of these vocalizations is decoded in a network of frontal and auditory brain regions. Neural decoding in the auditory regions, in particular, largely follows the acoustic properties that make whispered and voiced vocalizations considerably distinct on a perceptual level. Neural decoding also follows the general topographical frequency and temporal regularity maps in the auditory cortex. Finally, we found a hypercomplex and probably compensatory neural fronto-temporal network that supports the accurate affective classification of whispered vocalizations, which is probably based on their impoverished sound quality especially in terms of tonal information.

Author contributions

S.F. and D.G. designed the experiment. S.F. and W.T. acquired and analyzed the data. S.F., W.T. and D.G. wrote the manuscript.

Acknowledgments

This study was supported by the Swiss National Science Foundation (SNSF 105314_146559/1 and PP00P1_157409/1) and by the NCCR Affective Sciences at the University of Geneva (SNSF 51NF40-104897). We thank Luc Arnal and Anne-Lise Giraud for helpful comments on the manuscript. The authors declare to have no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.08.023.

References

Bachorowski, J.A., Owren, M.J., 2001. Not all laughs are alike: voiced but not unvoiced laughter readily elicits positive affect. Psychol. Sci. 12, 252–257.
Banse, R., Scherer, K.R., 1996. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 70, 614–636.
Bar, M., 2009. The proactive brain: memory for predictions. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 364, 1235–1243.
Beaucousin, V., Lacheret, A., Turbelin, M.R., Morel, M., Mazoyer, B., Tzourio-Mazoyer, N., 2007. FMRI study of emotional speech comprehension. Cereb. Cortex 17, 339–352.
Binder, J.R., Desai, R.H., Graves, W.W., Conant, L.L., 2009. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796.
Capilla, A., Belin, P., Gross, J., 2013. The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. Cereb. Cortex 23, 1388–1395.
Cirillo, J., Todt, D., 2002. Decoding Whispered Vocalizations: Relationships Between Social and Emotional Variables. pp. 1559–1563.
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. J. Neurosci. 31, 14067–14075.
Deshpande, G., LaConte, S., James, G.A., Peltier, S., Hu, X., 2009. Multivariate Granger causality analysis of fMRI data. Hum. Brain Mapp. 30, 1361–1373.
Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., Vuilleumier, P., 2012. Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. Cereb. Cortex 22, 191–200.
Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Gunther, M., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E., 2010. Multiplexed echo planar imaging for sub-second whole brain FMRI and fast diffusion imaging. PLoS One 5, e15710.
Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E., Dolan, R.J., 1997. Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6, 218–229.
Frühholz, S., Grandjean, D., 2012. Towards a fronto-temporal neural network for the decoding of angry vocal expressions. NeuroImage 62, 1658–1666.
Frühholz, S., Grandjean, D., 2013a. Multiple subregions in superior temporal cortex are differentially sensitive to vocal expressions: a quantitative meta-analysis. Neurosci. Biobehav. Rev. 37, 24–35.
Frühholz, S., Grandjean, D., 2013b. Processing of emotional vocalizations in bilateral inferior frontal cortex. Neurosci. Biobehav. Rev. 37, 2847–2855.
Frühholz, S., Fehr, T., Herrmann, M., 2009. Interference control during recognition of facial affect enhances the processing of expression specific properties—an event-related fMRI study. Brain Res. 1269, 143–157.
Frühholz, S., Ceravolo, L., Grandjean, D., 2012. Specific brain networks during explicit and implicit decoding of emotional prosody. Cereb. Cortex 22, 1107–1117.
Frühholz, S., Trost, W., Grandjean, D., 2014. The role of the medial temporal limbic system in processing emotions in voice and music. Prog. Neurobiol. 123, 1–17.

Frühholz, S., Gschwind, M., Grandjean, D., 2015a. Bilateral dorsal and ventral fiber pathways for the processing of affective prosody identified by probabilistic fiber tracking. NeuroImage 109, 27–34.
Frühholz, S., Hofstetter, C., Cristinzio, C., Saj, A., Seeck, M., Vuilleumier, P., Grandjean, D., 2015b. Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. Proc. Natl. Acad. Sci. U. S. A. 112, 1583–1588.
Frühholz, S., Trost, W., Kotz, S.A., 2016a. The sound of emotions—towards a unifying neural network perspective of affective sound processing. Neurosci. Biobehav. Rev.
Frühholz, S., van der Zwaag, W., Saenz, M., Belin, P., Schobert, A.K., Vuilleumier, P., Grandjean, D., 2016b. Neural decoding of discriminative auditory object features depends on their socio-affective valence. Soc. Cogn. Affect. Neurosci.
Giraud, A.L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 84, 1588–1598.
Glover, G.H., Li, T.Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44, 162–167.
Hackett, T.A., 2011. Information flow in the auditory cortical network. Hear. Res. 271, 133–146.
Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. "Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7, 213–223.
Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N.K., Uludag, K., Kayser, C., 2013. Spatial representations of temporal and spectral sound cues in human auditory cortex. Cortex 49, 2822–2833.
Hervais-Adelman, A.G., Carlyon, R.P., Johnsrude, I.S., Davis, M.H., 2012. Brain regions recruited for the effortful comprehension of noise-vocoded words. Lang. Cogn. Process. 27, 1145–1166.
Higashikawa, M., Nakai, K., Sakakura, A., Takahashi, H., 1996. Perceived pitch of whispered vowels-relationship with formant frequencies: a preliminary study. J. Voice 10, 155–158.
Jovicic, S.T., 1998. Formant feature differences between whispered and voiced sustained vowels. Acustica 84, 739–743.
Krause, J.C., Braida, L.D., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. J. Acoust. Soc. Am. 115, 362–378.
Ladich, F., 2007. Females whisper briefly during sex: context- and sex-specific differences in sounds made by croaking gouramis. Anim. Behav. 73, 379–387.
Lass, N.J., 1976. Speaker sex identification from voiced, whispered, and filtered isolated vowels. J. Acoust. Soc. Am. 59, 675.
Leitman, D.I., Wolf, D.H., Ragland, J.D., Laukka, P., Loughead, J., Valdez, J.N., Javitt, D.C., Turetsky, B.I., Gur, R.C., 2010. "It's not what you say, but how you say it": a reciprocal temporo-frontal network for affective prosody. Front. Hum. Neurosci. 4, 1–13.
Lewis, J.W., Talkington, W.J., Walker, N.A., Spirou, G.A., Jajosky, A., Frum, C., Brefczynski-Lewis, J.A., 2009. Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J. Neurosci. 29, 2283–2296.
Lewis, J.W., Talkington, W.J., Tallaksen, K.C., Frum, C.A., 2012. Auditory object salience: human cortical processing of non-biological action sounds and their acoustic signal attributes. Front. Syst. Neurosci. 6, 27.
Morrison, R., Reiss, D., 2013. Whisper-like behavior in a non-human primate. Zoo Biol. 32, 626–631.
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. J. Neurosci. 33, 19451–19469.
Norris, V., Cabin, A., Zemirline, A., 2005. Hypercomplexity. Acta Biotheor. 53, 313–330.
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, 767–776.
Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J. Neurosci. 24, 6810–6815.
Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. NeuroImage 119, 164–174.
Pulvermüller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., Shtyrov, Y., 2006. Motor cortex maps articulatory features of speech sounds. Proc. Natl. Acad. Sci. U. S. A. 103, 7865–7870.
Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. NeuroImage 52, 1059–1069.
Sato, J.R., Fujita, A., Cardoso, E.F., Thomaz, C.E., Brammer, M.J., Amaro, E., 2010. Analyzing the connectivity between regions of interest: an approach based on cluster Granger causality for fMRI data analysis. NeuroImage 52, 1444–1455.
Schwartz, M.F., 1967. Syllable duration in oral and whispered reading. J. Acoust. Soc. Am. 41, 1367–1369.
Seth, A.K., 2010. A MATLAB toolbox for Granger causal connectivity analysis. J. Neurosci. Methods 186, 262–273.
Seth, A.K., Barrett, A.B., Barnett, L., 2015. Granger causality analysis in neuroscience and neuroimaging. J. Neurosci. 35, 3293–3297.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Szameitat, D.P., Kreifelts, B., Alter, K., Szameitat, A.J., Sterr, A., Grodd, W., Wildgruber, D., 2010. It is not always tickling: distinct cerebral responses during perception of different laughter types. NeuroImage 53, 1264–1271.
Trost, W., Frühholz, S., Schon, D., Labbe, C., Pichon, S., Grandjean, D., Vuilleumier, P., 2014. Getting the beat: entrainment of brain activity by musical rhythm and pleasantness. NeuroImage 103, 55–64.
Trost, W., Frühholz, S., Cochrane, T., Cojan, Y., Vuilleumier, P., 2015. Temporal dynamics of musical emotions examined through intersubject synchrony of brain activity. Soc. Cogn. Affect. Neurosci. 10, 1705–1721.
Vestergaard, M.D., Patterson, R.D., 2009. Effects of voicing in the recognition of concurrent syllables. J. Acoust. Soc. Am. 126, 2860.
von Kriegstein, K., Smith, D.R.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30, 629–638.
Wiethoff, S., Wildgruber, D., Kreifelts, B., Becker, H., Herbert, C., Grodd, W., Ethofer, T., 2008. Cerebral processing of emotional prosody—influence of acoustic parameters and arousal. NeuroImage 39, 885–893.