Cerebral Cortex June 2013;23:1388–1395
doi:10.1093/cercor/bhs119
Advance Access publication May 17, 2012

The Early Spatio-Temporal Correlates and Task Independence of Cerebral Voice Processing Studied with MEG

Almudena Capilla[1,2], Pascal Belin[2,3,4] and Joachim Gross[2,3]

[1]Department of Biological and Health Psychology, Autonoma University of Madrid (UAM), Madrid, Spain, [2]Centre for Cognitive Neuroimaging (CCNi), Institute for Neuroscience and Psychology and [3]School of Psychology, University of Glasgow, Glasgow, UK and [4]International laboratories for Brain, Music and Sound (BRAMS), Université de Montréal & McGill University, Montreal, Quebec, Canada

Address correspondence to Pascal Belin, Institute for Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, G128QB Glasgow, UK. Email: pascal.belin@glasgow.ac.uk

Functional magnetic resonance imaging studies have repeatedly provided evidence for temporal voice areas (TVAs) with particular sensitivity to human voices along bilateral mid/anterior superior temporal sulci and superior temporal gyri (STS/STG). In contrast, electrophysiological studies of the spatio-temporal correlates of cerebral voice processing have yielded contradictory results, finding the earliest correlates either at ~300–400 ms, or earlier at ~200 ms ("fronto-temporal positivity to voice", FTPV). These contradictory results are likely the consequence of different stimulus sets and attentional demands. Here, we recorded magnetoencephalography activity while participants listened to diverse types of vocal and non-vocal sounds and performed different tasks varying in attentional demands. Our results confirm the existence of an early voice-preferential magnetic response (FTPVm, the magnetic counterpart of the FTPV) peaking at about 220 ms and distinguishing between vocal and non-vocal sounds as early as 150 ms after stimulus onset. The sources underlying the FTPVm were localized along bilateral mid-STS/STG, largely overlapping with the TVAs. The FTPVm was consistently observed across different stimulus subcategories, including speech and non-speech vocal sounds, and across different tasks. These results demonstrate the early, largely automatic recruitment of focal, voice-selective cerebral mechanisms with a time-course comparable to that of face processing.

Keywords: auditory cortex, human vocalizations, speech, superior temporal sulcus, temporal voice areas

Introduction

There is accumulating evidence for cortical areas with particular sensitivity to sounds of voice along the superior temporal sulci (STS) and superior temporal gyri (STG) of the human brain. Functional magnetic resonance imaging studies show that these "temporal voice areas" (TVAs) are more active in response to voices—speech or not—than to non-vocal sounds or control stimuli, such as scrambled voices and amplitude-modulated noise (Belin et al. 2000, 2002; Fecteau et al. 2004; von Kriegstein and Giraud 2004; Grandjean et al. 2005; Ethofer et al. 2009, 2012; Linden et al. 2011). TVAs are present in young infants (Grossmann et al. 2010; Blasi et al. 2011) as well as in the macaque brain (Petkov et al. 2008) where they contain "voice cells" highly selective for conspecific vocalizations (Perrodin et al. 2011), indicating a long evolutionary history and early development of cerebral voice processing.

Only few electroencephalography (EEG) studies have investigated the spatio-temporal correlates of cerebral voice processing, with divergent results. Measuring event-related potentials (ERPs) in response to sung voices and instruments, Levy et al. (2001, 2003) described a "voice-specific response" (VSR) peaking at around 320 ms after stimulus onset and dependant on the listener's attention. Charest et al. (2009) observed marked ERP differences on bilateral fronto-temporal electrodes when comparing a large set of vocal sounds—speech and non-speech—to non-vocal sound categories: this "fronto-temporal positivity to voice" (FTPV), maximal in the latency of the P200 component, emerged as early as 168 ms after stimulus onset. The FTPV has also been observed in 4- and 5-year-old children (Rogier et al. 2010). De Lucia et al. (2010) observed global field power (GFP) differences in a comparable time window (169–219 ms) when comparing human non-speech vocalizations to animal vocalizations. Source analyses attributed these GFP differences to a focal cluster along the right STS, close to, if not overlapping with, the TVAs; yet they were interpreted by the authors, somewhat surprisingly, as originating from the modulation in strength of a common cerebral network without topographic differences (De Lucia et al. 2010). That study also reported that responses to some subcategories of non-vocal sounds did not statistically differ from that to voice, questioning the selectivity of the effect.

These inconsistencies likely reflect differences in experimental protocol: studies used different sets of stimuli and different tasks. For instance, De Lucia et al. (2010)'s results relate to distractors in an oddball task, that is stimuli which the subject is actively ignoring and that are much more numerous than the target category, potentially reflecting uncontrolled attentional and adaptation effects.

We sought to clarify these apparently contradictory results by comparing, within a single experimental protocol, different subcategories of vocal and non-vocal sounds and different tasks varying in attentional demands. We also used magnetoencephalography (MEG) for a more accurate localization of underlying sources. We predicted that the earliest correlate of cerebral voice processing would be reflected in the FTPVm—the magnetic counterpart of the FTPV—with greater amplitude for vocal sounds irrespective of the listeners' task and with sources located close to the TVAs.

Materials and Methods

Participants

Ten healthy volunteers participated in the study (3 males; mean ± SD age, 31.2 ± 2.9). All participants were right handed and reported normal hearing. They all provided informed written consent and

received monetary compensation for their participation. The study was approved by the local ethics committee (University of Glasgow Faculty of Information and Mathematical Sciences) and conducted in conformity with the Declaration of Helsinki.

Stimuli and Task

Stimuli were selected in dependence of a behavioral study and some important acoustic properties. Firstly, in a behavioral study in which 10 volunteers participated, stimuli were selected out of a pool of 504 sounds: 252 "vocal" and 252 "non-vocal" stimuli (Belin et al. 2000). Vocal stimuli were defined as human vocalizations and consisted of either speech (e.g. vowel) or non-speech sounds (e.g. yawn). Non-vocal stimuli, defined as non-human vocalizations, represented the following 3 subcategories: animal vocalizations (e.g. horse), natural (e.g. waterfall), or artificial sounds (e.g. bell). Stimuli were edited using Cool Edit Pro (Syntrillium Corporation) to a sampling rate of 22 050 Hz and duration of 500 ms with a 10 ms linear attack and decay. They were root mean square (RMS) normalized using Matlab 7.5 (The MathWorks). The 504 stimuli of the pool were presented in random order, each sound being played twice. The inter-stimulus interval (ISI) ranged from 300 to 400 ms. Participants were asked to perform a forced-choice voice/non-voice categorization task. Stimuli were selected if they were recognized as either human vocal or non-vocal with correctness rates higher than 90%.

Secondly, stimuli were selected if they passed an intensity threshold of about 50% mean RMS intensity within the first 14 ms after stimulus onset. This selection step was aimed at avoiding potential jittering effects in the event-related fields (ERFs), previously observed in pilot MEG recordings. The intensity threshold was based on the pilot data. Overall, from the original pool of 504 stimuli, 115 vocal and 186 non-vocal stimuli remained. For the actual MEG experiment, we randomly selected 70 vocal and 70 non-vocal stimuli from them (with exception of vocal speech sounds that were entirely included as only 27 stimuli remained). The final set of stimuli consisted of 27 speech and 43 non-speech sounds; and 18 animal, 24 natural, and 28 artificial sounds. In order to test whether stimulus categories differed in acoustical features, we computed their average fundamental frequency (f0), standard deviation f0, and harmonic-to-noise ratio (HNR). Sixteen non-vocal stimuli were not included in the f0 analyses, as they did not have a clearly defined f0. Differences between stimulus types were statistically assessed via 2-sample *t*-tests.

The experimental task was presented using Psychtoolbox (Brainard 1997). Sound stimuli were delivered binaurally via a sound pressure transducer through two 5 m long plastic tubes terminating in plastic insert earpieces. Stimuli were presented at a self-adjusted comfortable level of about 65 dB SPL. During sound stimulation, a fixation cross was presented through a DLP projector (PT-D7700E-K, Panasonic) on a grey background at the center of the projection screen, subtending $0.9 \times 0.9°$ of visual angle. Participants were instructed to maintain their gaze at the fixation cross while listening to the sounds. The ISI varied randomly between 2 and 2.5 s. Sounds were presented in 50 trials blocks, comprising 25 vocal and 25 non-vocal randomly selected stimuli. The order of presentation was also randomized. In each block, participants were instructed to perform one of the following tasks: passive listening, a 1-back task or a categorization task. A brief reminder of the instructions of the current task was presented in the screen before the onset of each block. In the passive listening task, subjects were asked to simply listen to the sounds. In the 1-back task, they were instructed to press a non-magnetic response pad button (Lumitouch) with their right index finger whenever the current sound matched the previous one. The probability of match was 10%, and there were never 2 consecutive matches. In the forced-choice categorization task, participants were asked to press a button with their right index finger when they heard a human vocalization; and with their right middle finger when the sound was non-vocal (e.g. animal or environmental sounds). In the 2 tasks, subjects were instructed to delay their response until the offset of the auditory stimulation.

Participants went through 4 runs of the experimental task. They were given unlimited time to rest between runs. Each run consisted of 6 blocks that were pseudo-randomly assigned to the 3 tasks above described, each task being repeated twice. Blocks were interleaved with resting periods of 20 s, where participants were requested to remain still and relaxed. Each run comprised 300 trials, lasting approximately 15 min. In sum, participants performed 1200 trials, and each single stimulus was presented on average 8.5 times (8.5×140 stimuli $\approx$ 1200 trials). Before the experimental session, participants went through a practice session with 3 blocks, each assigned to each of the tasks. Practice blocks consisted of 10 trials and were interleaved with 10 s resting periods.

MEG Recording

Brain activity was recorded with a 248-magnetometers whole-head MEG system (MAGNES® 3600 WH, 4-D Neuroimaging) confined in a magnetically shielded room. MEG signal was acquired at a 508 Hz sampling rate and online high-pass filtered at 0.1 Hz.

Before starting the recording session, 5 coils were positioned on the participant's head, which was localized at the beginning and end of each run. These coils, together with 3 fiducial points and the subject's head shape, were digitized using a Polhemus system. During the recording session, subjects were seated in a reclining chair and supported their head against the back and top of the magnetometer. Participants were asked to remain as still as possible and were continuously monitored by a video camera. They were also instructed to minimize blinking during auditory stimulation. Eye movements were monitored using a SR remote Eyelink system (FL-890, SR Research Ltd.). Calibration of eye fixations was performed at the beginning of each run using a 9-point fixation procedure.

Eye-Movement Analysis

Analysis of eye-movement data aimed to test whether fixation was accurate and equivalent across different stimulus categories and tasks. One subject was excluded from this analysis due to recurrent eye-tracking signal loss. Fixation periods during stimuli presentation were extracted using in-house Matlab code (Matlab 7.5; The MathWorks). In order to determine whether subjects were significantly keeping fixation, we computed statistical maps of the number of fixations using the iMap toolbox (Caldara and Miellet 2011; http://perso.unifr.ch/roberto.caldara/index.php?page=4). Dispersion of the gaze was assessed in terms of standard deviation of gaze position, and compared across conditions by means of 2-sided paired samples *t*-tests.

MEG Analysis

The analysis of the MEG signal was performed using the FieldTrip software package (Oostenveld et al. 2011; http://www.ru.nl/fcdonders/fieldtrip/) and in-house Matlab code.

Preprocessing

The preprocessing of the MEG signal was carried out as follows. First, the signal was epoched in trials of 1 s length (200 ms pre-stimulus) time-locked to stimulus onset. Before visually inspecting MEG traces for artifacts, we removed the DC offset and linear trends in the signal to facilitate visualization. Three excessively noisy sensors were discarded from all subjects' analyses. Additionally, trials contaminated with eye blinks or squid jumps were discarded from further analysis. Then, signals recorded by the MEG reference sensors were used to reduce noise, as implemented in the "ft_denoise_pca" function in FieldTrip. The strongest component corresponding to the cardiac artifact was projected out of the MEG signal using independent component analysis (ICA) ("runica" algorithm implemented in FieldTrip/EEGLAB, http://sccn.ucsd.edu/eeglab/) after a dimensionality reduction to 20 components. Finally, the signal was digitally low-pass filtered below 30 Hz and baseline corrected using the 200 ms pre-stimulus time window.

Sensor-Level Analysis

Trials were split based on the stimulus category (vocal; non-vocal). Trials with incorrect responses in the categorization task were excluded from further analyses. Similarly, trials corresponding to both

false alarms and target stimuli in the 1-back task were also discarded. We then computed the ERFs elicited by each stimulus category for each subject and, subsequently, we computed the grand-average across participants. We identified the sensor exhibiting maximum activity and the peak latency in the grand-average ERF across conditions, and extracted the mean amplitude value in a 30 ms time window around the peak in the sensor of interest. We statistically tested differences between conditions in the 2 main components identified in the ERFs by means of 2-sided paired samples *t*-tests.

Source-Level Analysis

MEG–Magnetic Resonance Image Co-Registration. T1-weighted structural magnetic resonance images (MRIs) of each participant were co-registered to the MEG coordinate system by a semi-automatic procedure that provided the best fit between the subject's MRI and the digitized head shape. The scalp surface was extracted from the MRI by means of an erosion/dilation and thresholding procedure. To obtain a first approximate alignment between MEG and MRI coordinates, we manually located the 3 digitized fiducial points (nasion, left and right pre-auricular points) in the individual's MRI. Subsequently, we applied an iterative fitting procedure consisting of 2 steps. First, the rotation of the head shape in all directions in increasingly smaller rotation angles (±15°, ±7.5°, ±3.75°, ±1.87°, and ±0.94°), and second, the automatic fitting of head shape and scalp points by applying a modified version of the Iterative Closest Point algorithm (Besl and McKay 1992; A.S.Mian, icp2©: http://www.csse.uwa.edu.au/~ajmal/code/icp2.m). In each iteration, the relative position between head shape and scalp was updated to the one providing the minimum mean distance error.

Head and Forward Models. The brain surface was extracted from each MRI using the segmentation routine implemented in FieldTrip/SPM2 (http://www.fil.ion.ucl.ac.uk/spm). We then constructed a semi-realistic single shell head model (Nolte 2003) based on each individual's brain. Subsequently, we adapted a standard grid of 6 mm resolution derived from the Montreal Neurological Institute (MNI) brain to each subject's brain volume. This was achieved by normalizing the individual MRIs to the standard MNI brain through a linear affine transformation (FieldTrip/SPM2). The inverse of the resulting transformation matrix was applied to the MNI-standard grid to transform it into each subject's brain space. Finally, we computed and normalized the lead fields corresponding to the 2 tangential orientations for each voxel.

Source Localization Analysis. The localization of brain sources was performed by means of beamforming (Van Veen et al. 1997; Gross et al. 2001). We computed the covariance matrix for a time window ranging from 200 ms pre-stimulus to 400 ms post-stimulus. This covariance matrix was used to compute the spatial filter coefficients by means of linearly constrained minimum variance beamformer (Van Veen et al. 1997). In order to maximize the sensitivity of the beamformer to focal sources, we did not apply regularization. Subsequently, we projected the sensor-level signal of each trial into each voxel of source-space through a spatial filter corresponding to a dipole at this location with fixed optimal orientation. We then computed source-level ERFs for each stimulus subcategory (2 vocal and 3 non-vocal subcategories) and task (3 tasks), leading to 15 source-level ERFs per voxel and subject. The number of trials included in each ERF was made equal for all conditions, by randomly selecting N trials (where N corresponds to the minimum number of trials across conditions; mean 42.1 trials, range 39–46 trials). Thus, the number of trials actually used for computing the source-level ERFs was approximately 600 (42 trials × 3 tasks × 5 subcategories). On average, there were 4.5 presentations of each specific sound. The source-level ERFs were normalized as relative change with respect to the RMS of the baseline signal.

ICA on Source-Level Data. We performed ICA on the reconstructed source-level data in order to identify the brain patterns underlying the ERF components of interest (Fig. 1 for a schematic explanation). This methodological approach provided us information about both (i) the spatial signatures of voice processing as well as (ii) the time course of these brain patterns, which can be compared across stimulus categories. All the source-level ERFs extracted for each condition and subject (i.e. 15 × 10 ERFs for each grid voxel) were submitted to ICA ("runica"; FieldTrip/EEGLAB, http://sccn.ucsd.edu/eeglab/). Prior to ICA, we performed a principal component analysis (PCA)-based dimensionality reduction to 2 components corresponding to the 2 dominating features in the sensor-level signal (i.e. N1m and FTPVm). Final results did not substantially change when a higher number of dimensions were used (e.g. 3 or 6 dimensions). ICA analysis yielded 2 ICs, each of them characterized by a specific temporal course per condition and subject, as well as an underlying brain pattern, obtained by rectifying the IC spatial distribution. The IC time series were baseline corrected and combined accordingly to the differences between conditions to be tested in each case. As for the sensor-level analysis, we extracted the mean amplitude value along a time window of ±15 ms around the peak (the peak latency was identified based on the grand-average across conditions). Differences between the 2 sound categories (vocal and non-vocal) were statistically tested by means of 2-sided paired samples *t*-tests. In order to further investigate the stimulation and task specificity of the voice-sensitive ICs, we tested the activity elicited by the different stimulus subcategories as well as the consistency of vocal/non-vocal differences across tasks. As for the previous analysis, we computed the mean amplitude value in a 30 ms length time window around the

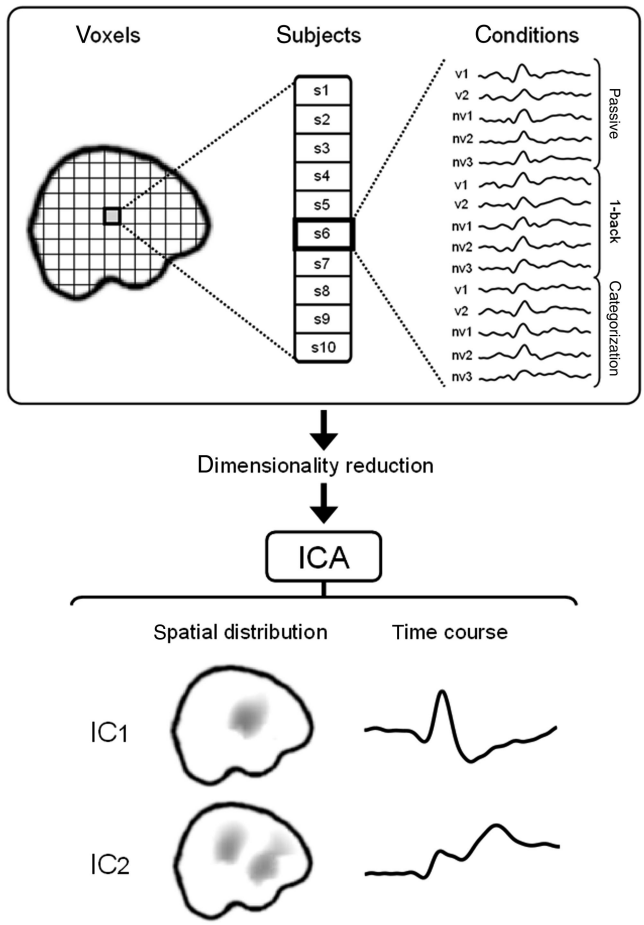

Figure 1. Schematic representation of the ICA analysis on source-level data. All the source-level ERFs previously computed for each voxel, subject and condition were submitted to a PCA-based dimensionality reduction and subsequent ICA. As a result, we obtained 2 ICs, each of them exhibiting a specific brain distribution and a time course per condition and subject. The figure shows the average time course collapsed across conditions and subjects. Abbreviations: vocal subcategories: v1, v2; non-vocal subcategories: nv1, nv2, nv3.

peak of interest, extracted from the grand-average over all conditions. We statistically tested the effect of stimulus subcategory by means of a one-way ANOVA. For the second test, we performed a 2-way ANOVA, with stimulus category (vocal/non-vocal) and task as factors.

Results

Behavioral Results

Behavioral results of the 1-back task showed a high level of accuracy, in terms of both hits (median 98%; interquartile range, IQR, 7%) and false alarms (median 0%; IQR 2%). Mean reaction time from stimulus offset for correctly detected targets was 400 ± 210 ms. Performance did not significantly differ between vocal and non-vocal stimuli in either hit rate, false alarm rate (Wilcoxon's signed-rank test, $P > 0.05$), or reaction time (2-sided paired samples *t*-test; $t_{(9)} = -1.53$, $P > 0.05$).

In the vocal/non-vocal categorization task, participants showed 97% (median; IQR 5%) of correct categorizations for vocal sounds and 98% (median; IQR 2%) for non-vocal sounds. Reaction times from stimulus offset were 520 ± 159 ms for vocal stimuli and 520 ± 161 ms for non-vocal stimuli. We did not observe differences in performance between both stimulus categories in either accuracy (Wilcoxon's signed-rank test, $P > 0.05$) or reaction time (2-sided paired samples *t*-test; $t_{(9)} = 0.21$, $P > 0.05$).

Fixation Accuracy and Equivalence Across Conditions

Statistical mapping of the number of fixations across subjects revealed 1 single significant cluster around the location of the fixation cross (maximum *z*-score = 20.6, $P < 0.001$ corrected for multiple comparisons). Pairwise comparisons of individual statistical fixation maps between conditions showed that maximum *z*-score did not differ between either stimulus categories (vocal vs. non-vocal stimuli: $t_{(8)} = 1.66$, $P > 0.05$) or tasks (passive vs. 1-back task: $t_{(8)} = 1.60$; passive vs. categorization task: $t_{(8)} = 0.56$; 1-back vs. categorization task: $t_{(8)} = -0.65$; $P > 0.05$). Gaze dispersion during stimulus presentation was 0.62 ± 0.53° (mean ± SD) for horizontal and 0.59 ± 0.28° for vertical eye fixation positions. The dispersion of the gaze did not significantly differ in either horizontal or vertical direction between sound categories ($t_{(8)} < 1.10$, $P > 0.05$) as well as between tasks ($t_{(8)} < 1.52$, $P > 0.05$).

Acoustical Differences Between Stimulus Categories

Average f0

Average f0 was 270 ± 92 Hz (mean ± SD) for vocal stimuli and 281 ± 163 Hz for non-vocal stimuli. Average f0 did not statistically differ between vocal and non-vocal stimuli ($t_{(122)} = -0.46$, $P = 0.646$).

Standard Deviation f0

Standard deviation f0 was 39.9 ± 37.9 Hz for vocal and 42.1 ± 52.5 Hz for non-vocal stimuli, and did not show statistical differences between stimulus categories ($t_{(122)} = -0.28$, $P = 0.782$).

Harmonic-to-Noise Ratio (HNR)

HNR was 13.5 ± 6.8 dB for vocal and 6.3 ± 8.8 dB for non-vocal stimuli. Vocal stimuli showed a higher HNR than non-vocal stimuli ($t_{(138)} = 5.40$, $P < 0.001$). In order to further explore the HNR differences across stimuli, we computed the HNR for the different stimulus subcategories. HNR was 17.7 ± 4.8 dB for speech stimuli, 10.8 ± 6.6 dB for non-speech sounds, 5.7 ± 4.8 dB for animal stimuli, −0.1 ± 2.3 dB for natural sounds, and 12.1 ± 10.3 dB for artificial sounds. Statistical analysis revealed that HNR differed between all the subcategories ($P < 0.05$) with the exception of the vocal non-speech and non-vocal artificial subcategories ($t_{(69)} = -0.65$, $P = 0.516$). In summary, HNR was different between stimulus subcategories within the same general category, and did not differ between non-speech and artificial sounds, each of them belonging to a different general category.

MEG Results: Sensor-Level Data

Figure 2 shows the ERFs elicited by auditory stimuli (collapsed across all conditions) (Fig. 2*A*), and the corresponding topographies (Fig. 2*B*) and synthetic planar gradient topographies (Fig. 2*C*) for the magnetic N1 (N1m), the magnetic "fronto-temporal positivity to voice" (FTPVm) and the sustained field (SF) components. The peak latency for the N1m component was 108 ms in both hemispheres. Peak latencies for the FTPVm were 216 ms for the left hemisphere (LH) and 226 ms for the right hemisphere (RH) sensors exhibiting maximum activity (sensors highlighted in Fig. 3).

We did not find differences between vocal and non-vocal stimuli in the amplitude of either left or right N1m ($t_{(9)} = 0.13$,

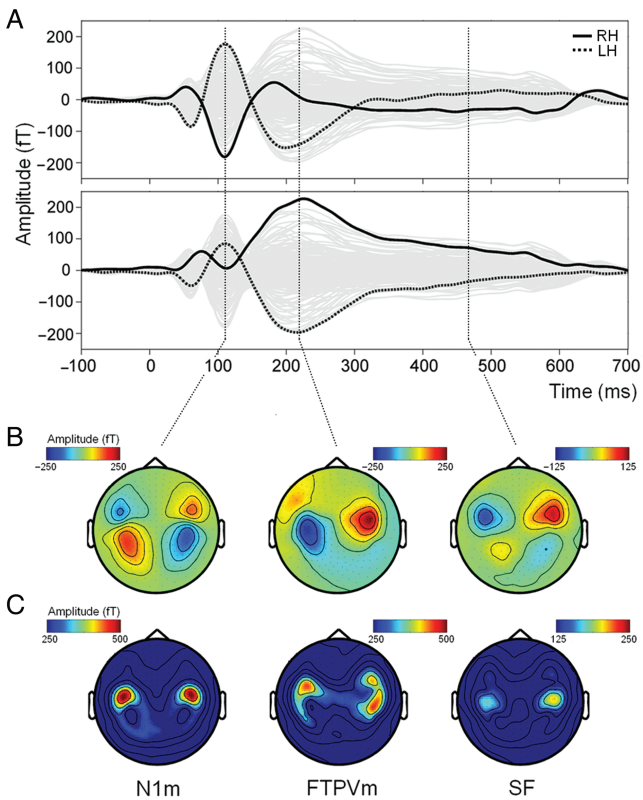

Figure 2. Sensor-level results: ERFs and topographies. (*A*) ERFs elicited by auditory stimuli collapsed across conditions. The light grey lines represent the ERFs of each MEG sensor. Black lines represent the channels exhibiting maximal activity for the N1m (upper panel) and FTPVm (lower panel) components in the RH (solid lines) and LH (dashed lines). (*B*) Topographies for the 3 main ERF components: the N1m at around 100 ms, the FTPVm peaking in the 200–250 ms time window, and the SF lasting for the duration of the auditory stimulation. (*C*) Synthetic planar gradients topographies for each ERF component, highlighting the sensors exhibiting the strongest gradients.

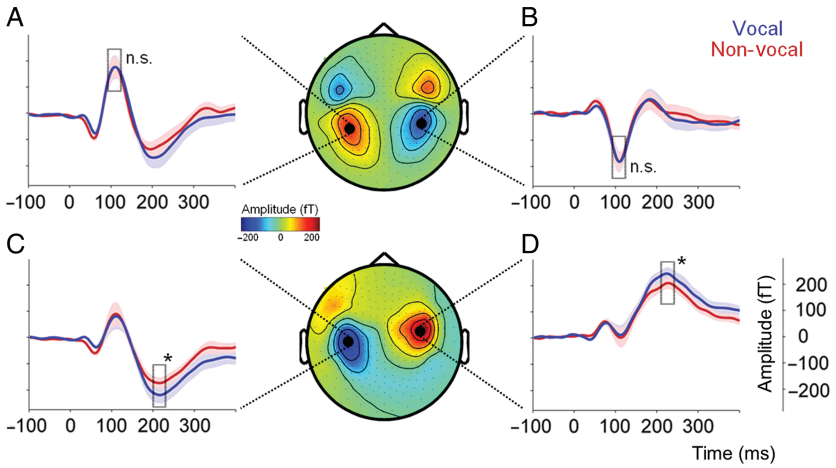

Figure 3. Sensor-level results: differences between vocal and non-vocal stimuli. ERFs of the sensors exhibiting maximum activity (indicated by black dots) for the (*A*) LH N1m, (*B*) RH N1m, (*C*) LH FTPVm, and (*D*) RH FTPVm components. ERFs evoked by vocal stimuli are represented by blue lines; ERFs elicited by non-vocal stimuli are depicted in red. Shaded areas indicate the standard error of the mean (SEM). Boxes around the ERF components of interest indicate the ±15 ms time window statistically tested (*$P < 0.05$; n. s., not significant).

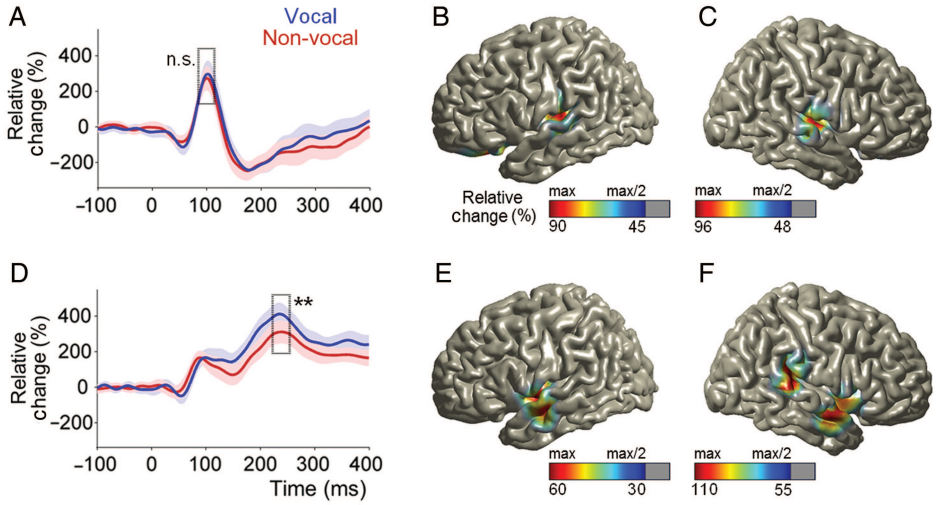

Figure 4. Source-level results. Independent component (IC) ERFs and brain distributions corresponding to the N1m and FTVPm. (*A*) N1m source-level ERF. (*B*) Brain pattern underlying the N1m IC, left view; (*C*) right view. (*D*) FTPVm source-level ERF. (*E*) Brain distribution of the FTPVm component, left view; (*F*) right view. Vocal stimuli are depicted in blue; non-vocal stimuli in red. Shaded areas indicate SEM; boxes indicate the ±15 ms time window around the peak of each component (**$P < 0.01$; n.s., not significant).

$t_{(9)} = -0.07$; $P > 0.05$) (Fig. 3*A* and *B*). However, FTPVm amplitude was significantly higher for vocal than non-vocal stimuli for both LH ($t_{(9)} = -2.96$, $P = 0.016$) and RH sensors ($t_{(9)} = 2.55$, $P = 0.031$) (Fig. 3*C* and *D*). The FTPVm component started to differ between vocal and non-vocal stimuli at 179 ms in the LH and at 181 ms in the RH ($t_{(9)} = -2.46$, $t_{(9)} = 2.44$, $P < 0.05$).

MEG Results: Source-Level Data

ICA revealed 2 ICs compatible with the sensor-level N1m and FTPVm components (Fig. 4*A* and *D*). The first IC was primarily correlated with the sensor-level N1m ($R^2 = 0.72$, $P < 0.001$), and the second IC with the sensor-level FTPVm component ($R^2 = 0.75$, $P < 0.001$). Additional analyses confirmed that these ICs were not substantially affected by changing the original data dimensionality reduction from 2 components to 3 or 6 components. Employing a PCA-based reduction to 3 components, the N1m and the FTPVm ICs were highly correlated with 2 corresponding ICs in both temporal course ($R^2 = 0.97$, $R^2 = 0.93$; $P < 0.001$) and brain distribution ($R^2 = 0.84$, $R^2 = 0.56$; $P < 0.001$). A dimensionality reduction to 6 components yielded also highly correlated ICs in time course ($R^2 = 0.61$, $R^2 = 0.87$; $P < 0.001$) as well as brain pattern ($R^2 = 0.69$, $R^2 = 0.73$; $P < 0.001$).

The peak latency for the source-level N1m was 100 ms. As for the sensor-level analyses, the N1m IC did not differ between vocal and non-vocal stimuli ($t_{(9)} = 1.00$, $P = 0.342$) (Fig. 4*A*). The local maxima of the spatial distribution of the source-level N1m component were located bilaterally along the middle extent of Heschl's gyrus in posterior superior temporal gyrus. The MNI coordinates of the peak voxel in the LH were [−42, −24, 16] ([*x*, *y*, *z*], Fig. 4*B*), and [54, −12, 10] in the RH (Fig. 4*C*).

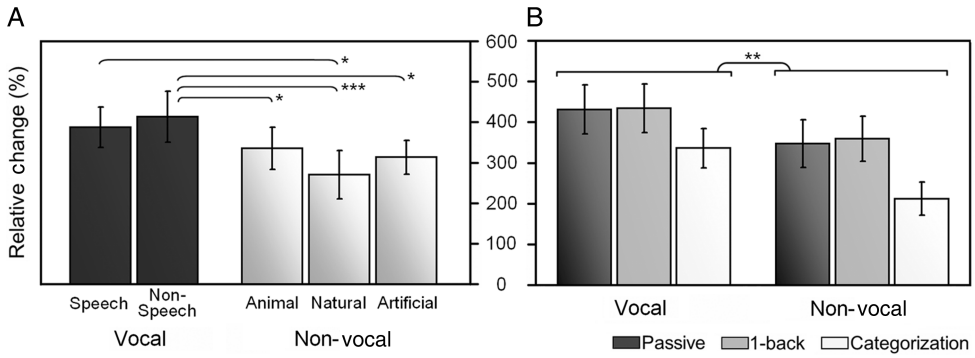

Figure 5. Subcategory and task × category effects. (*A*) Pairwise comparisons between stimulus subcategories (2 vocal subcategories: speech, non-speech; and 3 non-vocal subcategories: animal, natural, artificial). (*B*) Main effect of stimulus category (vocal vs. non-vocal). Error bars indicate ±SEM. Significance level: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

The IC corresponding to the FTPVm peaked at 238 ms. As for the sensor-level data, the amplitude of the FTPVm component was significantly higher for vocal than non-vocal stimuli ($t_{(9)} = 4.06$, $P = 0.003$) (Fig. 4*D*). The FTPVm component started to distinguish between stimulus categories at 147 ms ($t_{(9)} = 2.35$, $P < 0.05$). This component was bilaterally distributed over the mid-anterior part of the superior temporal sulcus, and along the planum temporale in the RH. The MNI coordinates of the peak voxel in the LH were [−48, −6, −2] (Fig. 4*E*). In the RH, the peak of maximum activity over mid-anterior STS was located at [54, 0, −10], and at [54, −30, 22] over posterior STG (Fig. 4*F*).

Subcategory Effect

The ANOVA revealed a main effect of stimulus subcategory ($F_{(4,36)} = 5.42$, $P = 0.002$). Subsequent post hoc analyses showed no significant differences within vocal or non-vocal subcategories, whereas most of the pairwise comparisons between vocal and non-vocal subcategories showed significant differences (speech vs. animal stimuli: $t_{(9)} = 1.46$, $P = 0.178$; speech vs. natural stimuli: $t_{(9)} = 2.67$, $P = 0.025$; speech vs. artificial stimuli: $t_{(9)} = 2.13$, $P = 0.062$; non-speech vs. animal stimuli: $t_{(9)} = 2.42$, $P = 0.039$; non-speech vs. natural stimuli: $t_{(9)} = 5.10$, $P < 0.001$; non-speech vs. artificial stimuli: $t_{(9)} = 2.69$, $P = 0.025$) (Fig. 5*A*).

Task × Category Effect

The task × category ANOVA showed, in line with the sensor-level results on FTPVm amplitude, a significant main effect of sound category ($F_{(1,9)} = 16.47$, $P = 0.003$) (Fig. 5*B*). The main effect of task did also result significant ($F_{(2,18)} = 13.25$, $P < 0.001$). In contrast, the interaction task × category did not show significant results ($F_{(2,18)} = 0.98$, $P = 0.396$), suggesting that differences between sound categories were consistent across tasks. In fact, pairwise comparisons showed significant, or close to significance, differences between vocal and non-vocal subcategories for all the tasks (passive listening task: $t_{(9)} = 4.13$, $P = 0.003$; 1-back task: $t_{(9)} = 2.17$, $P = 0.058$; categorization task: $t_{(9)} = 3.24$, $P = 0.010$).

Discussion

In this study, we employed MEG to investigate the earliest spatio-temporal correlates of cerebral voice processing. Our results reveal a preferential magnetic response for human vocalizations peaking in the 200–250 ms time range and emerging as early as 147 ms after stimulus onset, in good agreement with previous electrophysiological studies showing early correlates of voice processing (i.e. at ~200 ms) (Charest et al. 2009; De Lucia et al. 2010; Rogier et al. 2010). The timing of this response, as well as its sensitivity to vocal sounds, makes it suggestive of being the magnetic counterpart of the recently described FTPV (Charest et al. 2009; Rogier et al. 2010). The FTPVm originates from sources localized along the mid part of the STS bilaterally and in the right planum temporale, close if not overlapping with the TVAs (Belin et al. 2000). Its greater magnitude for vocal sounds was consistent across different stimulus subcategories and attentional demands.

The FTPVm: an Early Correlate of Voice Processing

In the present study, the magnetic response evoked by auditory stimuli was characterized by 3 components, corresponding to the N1m, the FTPVm, and the late SF. The FTPVm component showed a differential scalp distribution from the other 2 components (Fig. 2), suggestive of different generators as confirmed by source localization (Fig. 4).

In a previous MEG study, Gunji et al. (2003) found that although the N1m and the SF components were similarly localized around bilateral primary auditory cortex, only the SF was voice-sensitive. Here, we observed an earlier VSR, which might be due to differences in the materials employed. Gunji et al. (2003), as Levy et al. (2001, 2003), used sung voices from 4 singers compared with musical instruments sounds. Both studies observed differential activity between vocal and non-vocal stimuli in a relatively late time window (300–400 ms). However, the present study as well as recent EEG studies (Charest et al. 2009; De Lucia et al. 2010; Rogier et al. 2010) employed a larger set of stimuli that might have provided a better sampling of the vocal and non-vocal categories, thus contributing to the identification of earlier correlates of voice processing.

Additionally, the earlier and later correlates of voice processing might correspond to different stages in the perception of the voice. In analogy to Bruce and Young (1986)'s influential model of face perception, Belin et al. (2004) proposed that the information extracted in a first low-level auditory analysis could be subsequently used to encode a structural model of the voice. Vocal information might then be further analysed in interacting but partially segregated pathways for processing speech, affective, and identity information in parallel. EEG

studies on face processing have shown that face identification occurs in parallel and/or later than face discrimination (Eimer 2000; Schweinberger et al. 2002; Caharel et al. 2009, 2011). Thus, in analogy to the face perception literature, it might be suggested that the later correlates of voice processing found in studies using many repetitions of a limited number of speakers' voices might reflect the recognition of speaker's identity rather than voice discrimination per se (Levy et al. 2003). Further studies explicitly investigating the timing of different voice processing stages (e.g. Titova and Naatanen 2001; Beauchemin et al. 2006) might shed light on this issue.

Localization of the FTPVm Along the TVAs

Using source localization algorithms, De Lucia et al. (2010) localized the source of their observed vocal/non-vocal GFP differences to a focal cluster along right mid-STS—although they argued for a modulation in strength of a larger cerebral network without topographical differences. Here, we have exploited the greater localization power of MEG to demonstrate that the FTPVm's underlying sources are indeed focal and in very good correspondence with the anatomical location of the TVAs described by functional magnetic resonance imaging (fMRI) studies, including mid-anterior STS and right planum temporale (Belin et al. 2000, 2002; Fecteau et al. 2004; von Kriegstein and Giraud 2004; Grandjean et al. 2005; Ethofer et al. 2009, 2012; Latinus et al. 2011; Linden et al. 2011). Nevertheless, it should be noted that the lower spatial resolution of MEG when compared with fMRI might have precluded further separation of neighboring loci of activation along the STS previously identified in fMRI studies (Belin et al. 2000). Also, the high temporal resolution of MEG ensures that the TVA activation found here corresponds to the initial stages of voice perception, whereas fMRI activations might reflect a combination of early and late correlates of voice processing (e.g. reverberant activation).

In contrast to the right-lateralized activity found by De Lucia et al. (2010), we have found the early correlates of voice processing to be localized bilaterally, in agreement with previous fMRI studies. This discrepancy in the lateralization of voice-selective brain regions might be due to differences in the experimental design. Whereas De Lucia et al. (2010) compared non-speech human vocalizations with animal vocalizations, we employed a more extensive set of stimuli, that is 2 vocal and 3 non-vocal subcategories including those used by De Lucia et al. (2010). Additionally, they only examined distractors in an oddball task, that is stimuli being actively ignored by the subject and presented in a much larger proportion than the target stimuli, potentially contributing to uncontrolled top–down (attentional) and bottom–up (adaptation) effects. In contrast, we compared different tasks, which varied in attentional demands—but never involved selectively ignoring one category—and used vocal and non-vocal stimuli in equal proportions to avoid uncontrolled adaptation effects.

Consistency of the Results Across Stimulus Subcategories

The results of the comparisons between different stimulus subcategories are in line with our main result, that is higher activity in TVAs peaking at around 200–250 ms for vocal compared to non-vocal sounds. Furthermore, the comparisons between subcategories demonstrated the consistency of this finding, since (i) most of the single comparisons between subcategories replicated the general pattern found for the vocal/non-vocal categories, and (ii) none of the within-category comparisons revealed significant differences (Fig. 5*A*).

A particularly interesting finding refers to the speech/non-speech subcategories: De Lucia et al. (2010) suggested that voice-selectivity effects in Charest et al. (2009) might be confounded by the speech content of the stimuli—although they themselves did not compare speech and non-speech voice stimuli. The present results, as well as clear previous evidence (Belin et al. 2002; Charest et al. 2009), rule out this possibility as the preferential early response to vocal sounds in the TVAs was consistent across both speech and non-speech sounds.

Consistency of Voice-Selective Activity Across Different Attentional Demands

Levy et al. (2003) showed that the electrical VSR only distinguished between vocal and non-vocal stimuli if participants were selectively attending to the timbre of the sounds, leading to the conclusion that the VSR does not likely reflect an automatic response to voices. However, a large number of previous fMRI studies have identified voice-selective brain regions using passive listening tasks that do not require specific allocation of attention (Belin et al. 2000, 2002; Belin and Zatorre 2003; Fecteau et al. 2004). Similarly, previous EEG studies have identified earlier correlates of voice processing using tasks with diverse attentional demands, such as passive listening while watching a silent video (Rogier et al. 2010), detecting an occasional pure tone (Charest et al. 2009) or detection of an infrequent target category (De Lucia et al. 2010).

In the present study, we employed 3 different tasks to address this discrepancy, including passive listening, a 1-back task, and overt vocal/non-vocal categorization. Our results show that the differential brain response to human voices is largely consistent across tasks. Critically, the FTPVm: (i) is clearly present during passive listening, in line with fMRI studies; and (ii) remains consistent across different attentional demands (Fig. 5*B*) whether or not the subject's attention is focused on the vocal/non-vocal difference. Our results thus argue in favor of the automatic nature of the brain response elicited by human voices.

Influence of Low-Level Acoustical Features

Our results show that vocal and non-vocal stimuli did not differ in their fundamental frequency or its variation. In contrast, vocal stimuli were on average more harmonic (greater HNR) than non-vocal stimuli—a well-known characteristic of vocal sounds. This acoustical difference might partially contribute to the voice-selective brain response (Lewis et al. 2009; Leaver and Rauschecker 2010). Yet, we observed significant source activity differences between vocal and non-vocal sounds even for subcategories that did not differ in HNR—non-speech vocal sounds vs. artificial non-vocal sounds—showing that HNR differences do not account for all of the FTPVm.

Conclusion

In conclusion, we have identified the FTPVm—the magnetic counterpart of the FTPV—an early signature of cerebral voice processing located bilaterally along the TVAs. The sensitivity

of the FTPVm to vocal sounds is consistent across different subcategories and task demands. Our results are congruent with the view that low-level sound features are processed at the earlier latencies of the N1m component likely corresponding to activity close to the primary auditory cortex. Then, at approximately 200–250 ms, and emerging as early as 147 ms after stimulus onset, a more detailed structural analysis would allow for categorical distinctions between vocal and non-vocal stimuli. We suggest that this stage is indexed by the FTPV/m, which might be considered analogous to the face-sensitive N170/m component (Bentin et al. 1996).

Notes

We gratefully acknowledge the help of Frances Crabbe in data acquisition. We also thank Sébastien Miellet and Gavin Paterson for valuable advice regarding eye-tracking data analysis.

Funding

This work was supported by the UK's Economical and Social Research Council (ESRC) and Medical Research Council (MRC) (large grant RES-060-25-0010)

References

Beauchemin M, De Beaumont L, Vannasing P, Turcotte A, Arcand C, Belin P, Lassonde M. 2006. Electrophysiological markers of voice familiarity. Eur J Neurosci. 23:3081–3086.

Belin P, Fecteau S, Bedard C. 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn Sci. 8:129–135.

Belin P, Zatorre RJ. 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport. 14:2105–2109.

Belin P, Zatorre RJ, Ahad P. 2002. Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res. 13:17–26.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. 2000. Voice-selective areas in human auditory cortex. Nature. 403:309–312.

Bentin S, Allison T, Puce A, Perez E, McCarthy G. 1996. Electrophysiological studies of face perception in humans. J Cogn Neurosci. 8:551–565.

Besl P, McKay N. 1992. A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell. 14:239–256.

Blasi A, Mercure E, Lloyd-Fox S, Thomson A, Brammer M, Sauter D, Deeley Q, Barker GJ, Renvall V, Deoni S *et al*. 2011. Early specialization for voice and emotion processing in the infant brain. Curr Biol. 21:1220–1224.

Brainard DH. 1997. The psychophysics toolbox. Spat Vis. 10:433–436.

Bruce V, Young A. 1986. Understanding face recognition. Br J Psychol. 77(Pt 3):305–327.

Caharel S, d'Arripe O, Ramon M, Jacques C, Rossion B. 2009. Early adaptation to repeated unfamiliar faces across viewpoint changes in the right hemisphere: evidence from the N170 ERP component. Neuropsychologia. 47:639–643.

Caharel S, Jacques C, d'Arripe O, Ramon M, Rossion B. 2011. Early electrophysiological correlates of adaptation to personally familiar and unfamiliar faces across viewpoint changes. Brain Res. 1387:85–98.

Caldara R, Miellet S. 2011. iMap: a novel method for statistical fixation mapping of eye movement data. Behav Res Methods. 43:864–878.

Charest I, Pernet CR, Rousselet GA, Quinones I, Latinus M, Fillion-Bilodeau S, Chartrand JP, Belin P. 2009. Electrophysiological evidence for an early processing of human voices. BMC Neurosci. 10:127.

De Lucia M, Clarke S, Murray MM. 2010. A temporal hierarchy for conspecific vocalization discrimination in humans. J Neurosci. 30:11210–11221.

Eimer M. 2000. Event-related brain potentials distinguish processing stages involved in face perception and recognition. Clin Neurophysiol. 111:694–705.

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.

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P. 2009. Decoding of emotional information in voice-sensitive cortices. Curr Biol. 19:1028–1033.

Fecteau S, Armony JL, Joanette Y, Belin P. 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage. 23:840–848.

Grandjean D, Sander D, Pourtois G, Schwartz S, Seghier ML, Scherer KR, Vuilleumier P. 2005. The voices of wrath: brain responses to angry prosody in meaningless speech. Nat Neurosci. 8:145–146.

Gross J, Kujala J, Hamalainen M, Timmermann L, Schnitzler A, Salmelin R. 2001. Dynamic imaging of coherent sources: studying neural interactions in the human brain. Proc Natl Acad Sci USA. 98:694–699.

Grossmann T, Oberecker R, Koch SP, Friederici AD. 2010. The developmental origins of voice processing in the human brain. Neuron. 65:852–858.

Gunji A, Koyama S, Ishii R, Levy D, Okamoto H, Kakigi R, Pantev C. 2003. Magnetoencephalographic study of the cortical activity elicited by human voice. Neurosci Lett. 348:13–16.

Latinus M, Crabbe F, Belin P. 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex. 21:2820–2828.

Leaver AM, Rauschecker JP. 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J Neurosci. 30:7604–7612.

Levy DA, Granot R, Bentin S. 2001. Processing specificity for human voice stimuli: electrophysiological evidence. Neuroreport. 12:2653–2657.

Levy DA, Granot R, Bentin S. 2003. Neural sensitivity to human voices: ERP evidence of task and attentional influences. Psychophysiology. 40:291–305.

Lewis JW, Talkington WJ, Walker NA, Spirou GA, Jajosky A, Frum C, Brefczynski-Lewis JA. 2009. Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J Neurosci. 29:2283–2296.

Linden DE, Thornton K, Kuswanto CN, Johnston SJ, van de Ven V, Jackson MC. 2011. The brain's voices: comparing nonclinical auditory hallucinations and imagery. Cereb Cortex. 21:330–337.

Nolte G. 2003. The magnetic lead field theorem in the quasi-static approximation and its use for magnetoencephalography forward calculation in realistic volume conductors. Phys Med Biol. 48:3637–3652.

Oostenveld R, Fries P, Maris E, Schoffelen JM. 2011. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011:156869.

Perrodin C, Kayser C, Logothetis NK, Petkov CI. 2011. Voice cells in the primate temporal lobe. Curr Biol. 21:1408–1415.

Petkov CI, Kayser C, Steudel T, Whittingstall K, Augath M, Logothetis NK. 2008. A voice region in the monkey brain. Nat Neurosci. 11:367–374.

Rogier O, Roux S, Belin P, Bonnet-Brilhault F, Bruneau N. 2010. An electrophysiological correlate of voice processing in 4- to 5-year-old children. Int J Psychophysiol. 75:44–47.

Schweinberger SR, Pickering EC, Jentzsch I, Burton AM, Kaufmann JM. 2002. Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. Brain Res Cogn Brain Res. 14:398–409.

Titova N, Naatanen R. 2001. Preattentive voice discrimination by the human brain as indexed by the mismatch negativity. Neurosci Lett. 308:63–65.

Van Veen BD, van Drongelen W, Yuchtman M, Suzuki A. 1997. Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Eng. 44:867–880.

von Kriegstein K, Giraud AL. 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage. 22:948–955.