Cortical processing of pitch

NeuroImage 180 (2018) 291–300

Contents lists available at ScienceDirect

NeuroImage

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

Cortical processing of pitch: Model-based encoding and decoding of auditory fMRI responses to real-life sounds

Vittoria De Angelis [a,b], Federico De Martino [a,b,d], Michelle Moerel [c,a,b], Roberta Santoro [a,b], Lars Hausfeld [a,b], Elia Formisano [a,b,c,*]

[a] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, The Netherlands
[b] Maastricht Brain Imaging Center (MBIC), Maastricht University, Maastricht, The Netherlands
[c] Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, The Netherlands
[d] Center for Magnetic Resonance Research, University of Minnesota Medical School, 2021 Sixth Street SE, Minneapolis, MN 55455, United States

ARTICLE INFO

Keywords:
Auditory cortex
Pitch processing
Real-life sounds
fMRI
Decoding
Encoding

ABSTRACT

Pitch is a perceptual attribute related to the fundamental frequency (or periodicity) of a sound. So far, the cortical processing of pitch has been investigated mostly using synthetic sounds. However, the complex harmonic structure of natural sounds may require different mechanisms for the extraction and analysis of pitch. This study investigated the neural representation of pitch in human auditory cortex using model-based encoding and decoding analyses of high field (7 T) functional magnetic resonance imaging (fMRI) data collected while participants listened to a wide range of real-life sounds. Specifically, we modeled the fMRI responses as a function of the sounds' perceived pitch *height* and *salience* (related to the fundamental frequency and the harmonic structure respectively), which we estimated with a computational algorithm of pitch extraction (de Cheveigné and Kawahara, 2002). First, using single-voxel fMRI encoding, we identified a pitch-coding region in the antero-lateral Heschl's gyrus (HG) and adjacent superior temporal gyrus (STG). In these regions, the pitch representation model combining height and salience predicted the fMRI responses comparatively better than other models of acoustic processing and, in the right hemisphere, better than pitch representations based on height/salience alone. Second, we assessed with model-based decoding that multi-voxel response patterns of the identified regions are more informative of perceived pitch than the remainder of the auditory cortex. Further multivariate analyses showed that complementing a multi-resolution spectro-temporal sound representation with pitch produces a small but significant improvement to the decoding of complex sounds from fMRI response patterns.

In sum, this work extends model-based fMRI encoding and decoding methods - previously employed to examine the representation and processing of *acoustic* sound features in the human auditory system - to the representation and processing of a relevant perceptual attribute such as pitch. Taken together, the results of our model-based encoding and decoding analyses indicated that the pitch of complex real life sounds is extracted and processed in lateral HG/STG regions, at locations consistent with those indicated in several previous fMRI studies using synthetic sounds. Within these regions, pitch-related sound representations reflect the modulatory combination of height and the salience of the pitch percept.

Introduction

Pitch plays an essential role in auditory perception, enabling us, for example, to identify distinct speakers and to perceptually organize the acoustic elements of a complex scene (Bregman, 1990; Moore, 1995). For harmonic tones, pitch is the perceptual correlate of the fundamental frequency F0, that is the sound's lowest frequency value of which all the spectral components are an integer multiple. As the same pitch can be perceived even after removal of the energy at F0 (i.e. in the case of *missing fundamental*), pitch is more generally defined in relation to the repetition rate (or periodicity) of the temporal envelope of the sound. Indeed, the energy content at the fundamental frequency does not influence the periodicity of the temporal envelope, which is solely determined by the spacing of the harmonics (de Cheveigné, 2010).

The neural mechanisms underlying pitch perception are still largely debated. The “temporal” hypothesis assumes that the periodicity is

* Corresponding author. Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands.
E-mail address: e.formisano@maastrichtuniversity.nl (E. Formisano).

https://doi.org/10.1016/j.neuroimage.2017.11.020
Received 21 March 2017; Accepted 11 November 2017
Available online 13 November 2017

extracted based on the timing between successive spikes in the auditory nerve. In contrast, the "place" theory infers that pitch is determined by the harmonic template that best matches the spectral cues encoded *tonotopically* in the cochlea and throughout the ascending auditory pathway (Plack et al., 2005). Recent accounts suggest that both place and timing information are necessary in order to perceive the correct pitch (Oxenham, 2013; Oxenham et al., 2004; Shamma, 2004).

Several studies investigated the neural (fMRI) correlates of pitch processing in subcortical and cortical structures of the human auditory system by comparing the BOLD responses for a wide range of pitch evoking sounds and noise control stimuli. Using iterated ripple noise (IRN), Griffiths et al. (2001) found a positive correlation between temporal regularity and local brain activity in the cochlear nucleus (CN) and in the inferior colliculus (IC) bilaterally. Moreover, the contrast between time-varying and fixed pitch sequences revealed significant activation differences only in the auditory cortex, specifically in lateral Heschl's gyrus (HG) and in planum temporale (PT) bilaterally, as also revealed with PET (Griffiths et al., 1998). This suggested a hierarchy of pitch processing stages starting in the subcortical structures, which are sensitive to temporal regularity, and terminating at the cortical level, where perceived pitch (variations) are most likely encoded. Patterson et al. (2002) reported a selective activation in lateral HG both in response to pitch-producing IRN and melodic sounds. Barker et al. (2012) argued that the activity elicited by the IRN in lateral HG was due to the fine temporal structure of the stimuli instead of pitch *per se*, as the contrast between the responses to conventional IRN and "no pitch" IRN control sounds did not show a significant difference. A pitch-tuned region was identified in the "anterior half of the auditory cortex" in Norman--Haignere et al. (2013). The activation of this region was predominantly driven by the resolved harmonics of the stimuli, and overlapped with a low-frequency area in the tonotopy map. These results are consistent with single-unit recordings in marmoset monkeys, reporting pitch-selective neurons located in a low-frequency region near the antero-lateral border of the primary auditory cortex (Bendor and Wang, 2005), potentially corresponding to lateral HG in humans (Bendor, 2012; Bendor and Wang, 2006). In addition, selective activation in response to pitch-evoking dichotic stimuli (Huggins pitch) has been observed in PT (Garcia et al., 2010; Hall and Plack, 2007, 2009). Importantly, a covariation of neural activity and pitch salience (dissociated from the physical stimulus regularity) was revealed in a cortical area located in the antero-lateral end of HG bilaterally (Penagos et al., 2004), whereas no such relation has been found for the PT region. In summary, fMRI findings support the hypothesis that the auditory cortex is involved in pitch perception. However, the exact location of a presumed pitch processing center in the human auditory cortex remains controversial (Griffiths and Hall, 2012).

The above-mentioned studies examined pitch processing by measuring fMRI responses to synthetic stimuli. However, for sounds occurring in everyday life pitch perception is more complex than for these artificial stimuli. For instance, the pitch of complex sounds may be influenced by the sound's overall spectral content and especially by the spectral locus of maximum energy concentration, which also relates to the brightness of timbre (de Cheveigné, 2005). Moreover, the strength of the pitch percept (or *salience*) is influenced by the degree to which the spectral components of sounds are harmonic, such that inharmonic sounds tend to evoke a pitch less salient than the one evoked by harmonic tones (Houtsma, 1997). As most of the sounds originating from natural and man-made sources are not perfectly harmonic, the brain processing underlying pitch perception for real-life sounds necessarily entails computational and representational mechanisms for extracting and combining multiple dimensions of pitch, notably pitch height (i.e. the dimension of pitch specifically related to F0) and pitch salience (i.e. the dimension of pitch related to sound harmonic structure).

The aim of the present study was to investigate these mechanisms in human auditory cortex through the model-based analysis of 7 T fMRI responses to real life sounds. First, we used single-voxel encoding (Kay et al., 2008b) and modeled the fMRI responses to a large set of complex naturalistic sounds as a function of sound representation models incorporating information on pitch *height* alone, pitch *salience* alone or on a *weighted* combination of height and salience. Both pitch height and salience were estimated using the YIN algorithm (de Cheveigné and Kawahara, 2002). Then, we evaluated the capability of these various models to predict the responses to a left-out sample of stimuli. The prediction accuracy obtained for these models were compared to each other and to the accuracy obtained with models describing the sounds by their spectral energy content on the same set of features as the pitch models (i.e. frequency bins). Results showed that fMRI responses in cortical regions located bilaterally in lateral HG and adjacent STG were predicted better by the pitch-based than by the energy-based sound representations. Moreover, in the right hemisphere regions, the prediction accuracy for the model combining pitch height and salience was significantly better than for the other pitch models.

Our previous work has shown that fMRI single-voxel responses (Santoro et al., 2014) and response patterns (Santoro et al., 2017) to natural sounds can be predicted accurately by a sound representation model based on the combination of spectro-temporal modulations (Chi et al., 2005). Sound representations explicitly encoding for pitch are expected to provide complementary and relatively independent information on the sound. In fact, current models of auditory scene analysis hypothesize that the auditory system uses pitch in parallel to the multi-resolution representation for parsing the auditory objects of complex scenes (Elhilali and Shamma, 2008; Shamma et al., 2011). Thus, a final aim of the study was to test whether a sound representation model based on pitch - used as a complement to the multi-resolution model - can provide additional information for decoding complex sounds from fMRI response patterns. We addressed this question using model-based multi-voxel decoding (Miyawaki et al., 2008; Santoro et al., 2017). Results showed that pitch information contributed to sound decoding significantly only for circumstantiated regions in lateral HG and STG and not in the remainder of the auditory cortex, which supports the hypotheses on the relevance of these regions for coding pitch information.

Materials and methods

Subjects and ethical statement

Five healthy subjects that were different for the two experiments participated in Experiment 1 ($n_1 = 5$, median age = 32, three males) and Experiment 2 ($n_2 = 5$, median age = 27 years, two males). The data of Experiment 1 and Experiment 2 have been previously described (Exp. 1: De Martino et al., 2013; Moerel et al., 2013; Santoro et al., 2014; Exp. 2: Santoro et al., 2017, publicly available at https://doi.org/10.5061/dryad.np4hs) and are analyzed here using a new approach. In this section the relevant elements of experimental procedures and fMRI response estimation will be described. All subjects (Experiment 1 and Experiment 2) reported no history of hearing disorder or neurological disease, and gave informed consent before commencement of the measurements. The Institutional Review Board for human subject research at the University of Minnesota (Experiment 1) and the Ethical Committee of the Faculty of Psychology and Neuroscience at Maastricht University (Experiment 2) granted approval for the study. Procedures followed the principles expressed in the Declaration of Helsinki. Informed consent was obtained from each participant before conducting the experiments.

Experimental procedures and fMRI responses estimation

Stimuli consisted of recordings of natural sounds including speech, voices, animal cries, scenes from nature, musical instruments and tool sounds (168 and 288 sounds for Experiment 1 and 2 respectively, 16 000 Hz sampling frequency, 1000 ms duration). In Experiment 1, for each subject 8 functional runs were collected; 144 sounds were presented in 6 training runs with 3 repetitions overall while the remaining 24

sounds were presented in 2 testing runs with 3 repetitions per run. In Experiment 2, the stimuli were divided in 4 non-overlapping sets (72 sounds each) that equally represented the semantic categories (i.e., 12 sounds per semantic category). The subjects underwent two scanning sessions, each consisting of 6 runs. Per session, 2 sound sets were repeated in 3 runs and each stimulus was presented 3 times across runs. In both experiments fMRI time series were acquired according to a fast event-related design (Experiment 1: TR = 2600 ms; TA = 1200 ms; TE = 30 ms; GRAPPA acceleration X3; partial Fourier 6/8; voxel size = 1.5 × 1.5 × 1.5 mm^3; silent gap = 1400 ms; Experiment 2: TR = 2600 ms; TA = 1200 ms; TE = 19 ms; GRAPPA acceleration X2; partial Fourier 6/8; voxel size = 1.5 × 1.5 × 1.5 mm^3; silent gap = 1400 ms). Sounds were presented in the silent gap between acquisitions with a randomly assigned inter-stimulus interval of 2, 3 or 4 TRs. The data was preprocessed with BrainVoyager QX (Brain Innovation, Maastricht, the Netherlands; temporal high pass filter, and 3D motion correction) and sampled in Talairach space. Next, for each voxel, the hemodynamic response function (HRF) common to all stimuli was estimated via a deconvolution analysis in which all stimuli were treated as a single condition. The fMRI responses to the stimuli (which will be referred to as "beta weights") were then computed by using the estimated HRF with one predictor per sound (Kay et al., 2008a). In Experiment 2 a 4-fold cross validation across the 4 stimulus sets was implemented. In both experiments, the HRF was estimated using the training data and beta weights were computed separately for training and testing sounds. Further analyses were performed on voxels with a significant positive response to the training sounds ($p < 0.05$, uncorrected) within an anatomically defined mask, which included Heschl's gyrus (HG), planum polare (PP), planum temporale (PT), and superior temporal gyrus (STG).

Pitch and sound representation models

We considered four different sound representation models: 1) a *Weighted Pitch* model, representing the perceived pitch of natural sounds as a "weighted" combination of F0 contour (pitch height) and salience (Fig. 1A); 2) a *Pitch* model, representing the perceived pitch as the pitch height alone (Fig. 1B); 3) a *Tonotopy* model, which described each stimulus by its spectral energy (Fig. 1C) and 4) a *Timbral Brightness* model, representing the sounds by the height of the spectral centroid (Fig. 1D), which is perceptually related to the brightness of timbre. All the models were implemented with custom Matlab (The MathWorks Inc.) code using the same time-frequency resolution.

Pitch and Weighted Pitch models

The perceived pitch of each sound was modeled based on the fundamental frequency estimated with the YIN algorithm (de Cheveigné and Kawahara, 2002). The algorithm detects the periodicity (and thus the fundamental frequency) of a given signal by measuring its self-similarity across time through the following *cumulative mean normalized difference function*:

$$d_t'(\tau) = \begin{cases} 1, & \text{if } \tau = 0 \\ \dfrac{d_t(\tau)}{(1/\tau)\sum_{j=1}^{\tau} d_t(j)}, & \text{otherwise;} \end{cases} \quad (1)$$

where $d_t(\tau) = r_t(0) + r_{t+\tau}(0) - 2r_t(\tau)$ and $r_t(\tau)$ is the autocorrelation function (ACF). For a periodic sound, the difference function in Eq. (1) is zero at each time lag τ integer multiple of the fundamental period. For sounds that are not perfectly periodic, the fundamental frequency is determined by the first time lag corresponding to a local minimum. Values of the difference function reflect the aperiodic (or inharmonic) component.

We applied the YIN algorithm on our stimuli, and derived two pitch representation models from the algorithm's output in the following way.

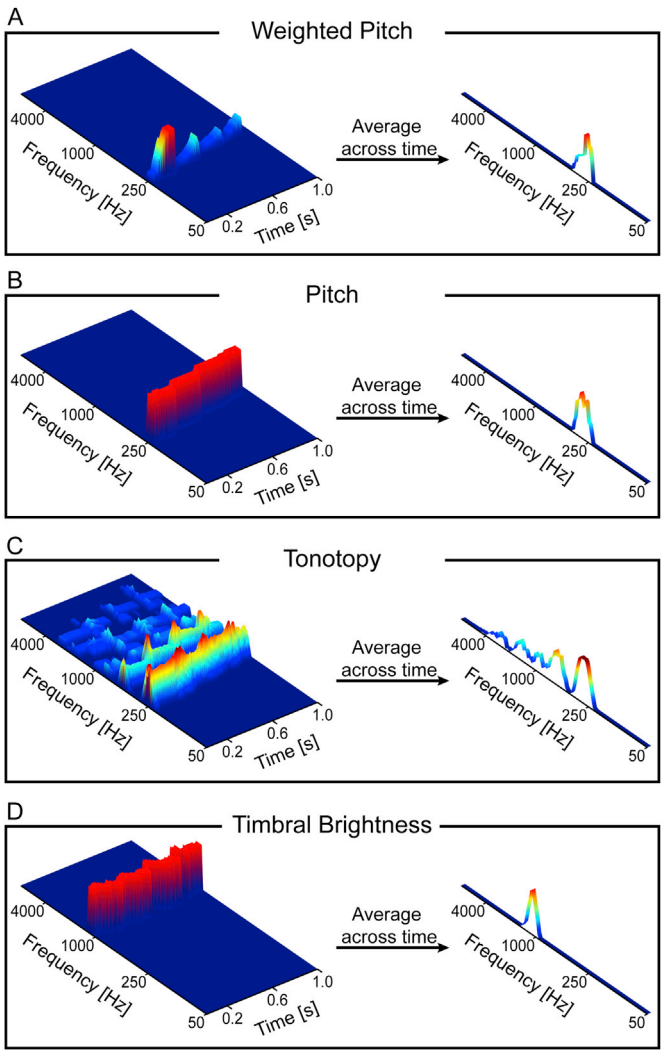

Fig. 1. Overview of the main stimulus representation models. (A) *Weighted Pitch* model: the perceived pitch is modeled as a "weighted" combination of pitch height and salience, obtained as a point by point multiplication between the time-resolved F0 contour and the coefficient reflecting the harmonic structure of the stimulus. This operation embeds the hypothesis that the information about the harmonic structure of the stimuli (perceptual salience) contributes in explaining the pitch-related cortical activity. After averaging across time, each stimulus is represented as a function of frequency, whose peak location corresponds to the estimated pitch height. Salience only influences the amplitude of the peak (related to pitch strength). (B) *Pitch* model: salience is not taken in account and pitch is modeled only by the F0 contour. (C) *Tonotopy* model: each sound is described by the spectral energy of each frequency component computed with the STFT. (D) *Timbral Brightness* model: stimuli are represented by the height of the spectral centroid, perceptually related to the brightness of timbre.

The time-resolved fundamental frequency $f_0(t)$ estimated by the YIN algorithm was discretized by defining a set of K = 128 logarithmically spaced frequencies F (from 50 to 8000 Hz) and selecting the k-th frequency as follows:

$$\mathrm{F0}(t) = \underset{\mathbf{F}}{\operatorname{argmin}}(|\mathrm{F}_k - f_0(t)|). \quad (2)$$

The resulting 2D representation was then averaged over time, obtaining a *Pitch* model which characterized the sounds by the averaged F0 contour (pitch height) (Fig. 1B).

Additionally, we derived a second model of pitch representation (referred to as a *Weighted Pitch* model) that characterized the sounds by both the estimated pitch height and the perceptual strength (or salience).

Specifically, we considered the difference function in Eq. (1) as a confidence indicator of the estimated F0 (de Cheveigné and Kawahara, 2002) and assumed that it reflects the perceived salience of the pitch. This assumption is in line with previous studies that considered the amplitude of the autocorrelation function as a quantitative descriptor of the strength of the perceived pitch (Leaver and Rauschecker, 2010; Meddis and Hewitt, 1991; Patterson et al., 1996; Yost et al., 1996). We obtained a coefficient reflecting the salience of the corresponding pitch as the inverse of $d_t'(\tau)$ normalized to the highest value in the whole set of sounds in each experiment. Low values corresponded to a low reliability of the estimated F0 value, representing a less salient pitch. The 2D representation of the F0 contour was then "weighted" by the reliability coefficient through a point-by-point multiplication and averaged across time (*Weighted Pitch* model, Fig. 1A). Additionally, a logarithmic transformation was applied to the resulting feature vector **wp** to balance the differences in order of magnitude of the confidence indicator:

$$\mathbf{wp}_l = \frac{1}{1 - \log_{10}(\mathbf{wp})}. \tag{3}$$

Note that the *Weighted Pitch* and *Pitch* models only differed by the weighting operation, which embedded the hypothesis that the information about the harmonic structure of the stimuli (perceptual salience) might contribute to the fMRI activity related to pitch.

To exclude the possibility that the observed differences between the *Weighted Pitch* and the *Pitch* model (see Results) depended on the overall difference of fMRI response levels to sounds with high vs low periodicity strength values, we included two additional models referred to as *Saliency* and *Saliency-Pitch* models respectively. The *Saliency* model represented sounds by the strength of the corresponding pitch (perceptual salience) as estimated by averaging the feature vector of the *Weighted Pitch* model (K = 1 feature). Note that pitch salience is closely related to the harmonic-to-noise ratio (HNR) (Giordano et al., 2013; Leaver and Rauschecker, 2010; Lewis et al., 2009). The *Saliency-Pitch* model, instead, was based on a combination of pitch strength and height which differed from that of the *Weighted Pitch* model. More specifically, the sound feature estimated with the *Saliency* model was appended to the feature vector of the *Pitch* model (K = 129 features).

Tonotopy and Timbral Brightness models

We compared the described pitch models to a *Tonotopy* model that represents the stimuli by their spectral content and a *Timbral Brightness* model, which reflects a perceptual property of complex sounds (de Cheveigné, 2005) and provides a sparse sound representation similarly to the pitch models.

The *Tonotopy* model was obtained by calculating the time-frequency representation of the acoustic energy with the Short Time Fourier Transform (STFT). The resulting spectrogram was downsampled to K = 128 logarithmically spaced frequencies between 50 and 8000 Hz and averaged across time (Fig. 1C). The *Timbral Brightness* model was obtained by computing the spectral centroid *SC(t)* as the power distribution over frequency at time *t* and selecting the *k*-th frequency as follows:

$$sc(t) = \frac{\sum_k \mathrm{F}_k \mathrm{X}_k(t)}{\sum_k \mathrm{X}_k(t)}; \tag{4}$$

$$\mathrm{SC}(t) = \underset{\mathbf{F}}{\operatorname{argmin}}(|\mathrm{F}_k - sc(t)|); \tag{5}$$

where X_k denotes the amplitude of each harmonic and F_k is the corresponding frequency value. The 2D representation of the time-resolved spectral centroid was then averaged across time (Fig. 1D). Note that here we defined *Timbral Brightness* as the center of gravity of the power spectrum, estimated as the weighted sum of frequencies. However, other measures are possible for the spectral centroid (Kendall et al., 1999; Marozeau et al., 2003; McAdams et al., 1995).

Single-voxel encoding

Estimation of the predicted fMRI responses

For each of the sound representation models described above, we derived the representations of all the stimuli, obtaining an (S × K) feature matrix **F**, where S is the number of sounds and K is the number of features. The fMRI responses $\boldsymbol{y}_i = [y_{1i}, \ldots, y_{Si}]^{\mathrm{T}}$ of the *i*-th voxel were then expressed as a linear transformation of the model features **F**:

$$\mathbf{y}_{\mathrm{S_{train}},i} = \mathbf{F}_{\mathrm{S_{train}}}\mathbf{w}_i + \mathbf{n}_i; \tag{6}$$

where $\mathrm{S_{train}}$ is the set of training sounds and **n** is a noise term. The overall voxel feature profile $\mathbf{w}_\mathrm{i} = [w_{1i}, \ldots, w_{\mathrm{K}i}]^{\mathrm{T}}$ (w_{ki} is the contribution of the *k*-th feature) was computed solving Eq. (6) with kernel ridge regression (Bishop, 2006; Hoerl and Kennard, 1970). The regularization parameter was determined independently for each voxel by generalized cross validation (Golub et al., 1979). Responses to the testing sounds were then predicted using the estimated regression weights:

$$\hat{\mathbf{y}}_{\mathrm{S_{test}},i} = \mathbf{F}_{\mathrm{S_{test}}}\mathbf{w}_i. \tag{7}$$

FMRI responses and features in the training data were normalized by removing the mean and dividing by the standard deviation across stimuli. The mean and standard deviation of the training data were used to normalize the test data (fMRI responses and features). In Experiment 2 we implemented a 4-fold cross validation scheme across the 4 non-overlapping stimulus sets (see above) and the procedure was repeated independently for each cross validation.

Voxel-based model comparison

Voxels' prediction accuracy was defined as the voxel-wise Pearson's correlation coefficient between measured and predicted fMRI responses to the testing stimuli. For each subject in Experiment 2 accuracy was averaged across the cross validations (Fisher transform/inverse transform was applied before/after the average). Single-subject maps of accuracy values were projected and smoothed (filter width = 4 vertices) on subject-specific cortical surfaces. Individual cortical surfaces and corresponding maps were aligned across subjects using Cortex Based Alignment (CBA) (Goebel et al., 2006). Group maps of model fit were obtained by color-coding the median value of vertices that had been included in the analysis of at least 8 out of the 10 subjects. Significance of the contrast between two models was assessed by performing a group random-effects non parametric test on the Fisher transform of the correlation values. The test statistic was defined for each vertex *i* as the group average of the individual difference $\mathbf{d}_i = \mathbf{r}_{i,\mathrm{Model}_1} - \mathbf{r}_{i,\mathrm{Model}_2}$. We computed the null distribution by changing the sign for a randomly selected subset of subjects and re-computing the test statistic. This procedure was repeated for all possible permutations of sign change (2^{N}) and the *p* value was computed as the proportion of values in the null distribution equal or higher than the observed average difference. Data of the two hemispheres were pooled together and a cluster size threshold procedure was performed (Forman et al., 1995). The cluster-level false-positive rate was estimated for each permutation using an initial vertex-level threshold set to $p = 0.05$. The minimum cluster size threshold which yielded a cluster-level false-positive rate (alpha) of 5% was then applied to the statistical maps.

Automated definition of the Pitch ROI and multi-voxel decoding

The described voxel-based model comparison relies on the spatial realignment of anatomical/functional data across subjects. To verify the consistency of the results across subjects, we performed an additional analysis aimed at identifying in each individual participant, a *Pitch ROI*. Furthermore, for these *Pitch ROIs*, a model-based multivariate decoding analysis was conducted as a complementary analysis to single voxel

encoding and to assess the hypothesis that the information about perceived pitch is represented preferentially within the selected region.

Using the training data only (Eq. (6)), a *Pitch ROI* (PR) was defined as the set of voxels for which the *Weighted Pitch* model fit was most significant ($p \leq 0.005$, uncorrected). The significance level was computed based on the comparison of the actual fit with the model fits obtained after permuting (200 permutations) the stimulus labels. Both the maps of actual and permuted model fit were spatially smoothed in the 3D volume space (Gaussian kernel, 3 mm FWHM), independently for each permutation (and cross validation in Experiment 2). This procedure was repeated for each single subject (Fig. S1).

For these individually determined ROIs we conducted a model-based decoding analysis aimed at reconstructing the model features from fMRI response patterns. The feature matrix **F** consisting of the representations of all the stimuli obtained for the *Weighted Pitch* model (see above) was expressed as a linear transformation of the multivoxel pattern response **Y** plus a bias term b and a noise term n:

$$\mathbf{f}_{S_{train},k} = \mathbf{Y}_{S_{train}}\mathbf{w}_k^T + b_k 1 + \mathbf{n}_k; \quad (8)$$

where S_{train} is the set of training sounds and **1** is an all-ones vector. Voxels' contribution to the k-th feature $\mathbf{f}_k$ ($\mathbf{w}_k = [w_{k1}, \ldots, w_{kI}]$, I = number of voxels) was computed solving Eq. (8) with kernel ridge regression (Bishop, 2006; Hoerl and Kennard, 1970) and the regularization parameter was determined independently for each feature by generalized cross validation (Golub et al., 1979). Features in the testing sounds were then reconstructed using the estimated regression weights as follows:

$$\widehat{\mathbf{f}}_{S_{test},k} = \mathbf{Y}_{S_{test}}\mathbf{w}_k^T. \quad (9)$$

The overall performance of the model was quantified performing a sound identification analysis on the basis of all reconstructed features. For each testing sound s, we computed Pearson's correlation coefficient (r_s) between the set of original and of reconstructed features. The normalized rank m of the correlation was used as a measure of the ability to correctly identify each sound:

$$m_s = 1 - \frac{\mathrm{rank}(r_s) - 1}{S_{test} - 1}. \quad (10)$$

A final identification accuracy per subject was then obtained as the average of m_s across sounds (and cross validations in Experiment 2).

The same analyses were performed separately on a control region consisting of all the remaining voxels not included in the PR, which was referred to as *Complementary Pitch ROI* ($\overline{\mathrm{PR}}$). The identification accuracies from *Pitch ROI* and *Complementary Pitch ROI* were compared performing a paired t-test on the Fisher transform of the accuracy values.

Combining pitch and spectro-temporal modulations for multi-voxel decoding

Previously, we have shown that fMRI responses to natural sounds can be predicted accurately by a sound representation model based on the combination of spectro-temporal modulations (Santoro et al., 2014, 2017). Thus, here we examined whether the *Weighted Pitch* model (i.e. the best performing of the pitch models tested, see Results) contributed relevantly to the fMRI-based decoding of sounds in addition to a multi-resolution modulation-based sound representation model.

First, we estimated a modulation-based representation of each stimulus (*Modulation* model) by applying the cortical stage of the "NSL Tools" package (available at http://www.isr.umd.edu/Labs/NSL/Software.htm) to the spectrogram obtained with the STFT. This cortical stage consists of a bank of 2D modulation selective filters tuned to spectral modulation frequencies of $\Omega = [0.5, 1, 2, 4]$ cyc/oct and temporal modulation frequencies of $\omega = [1, 3, 9, 27]$ Hz. The filter bank output was computed at each frequency along the tonotopic axis and then averaged over time. In order to decode the same number of features as for the *Weighted Pitch* model, we reduced the number of frequency bins to 8 (with constant bandwidths) and averaged the modulation energy within each of these bins. This resulted in K = 128 features in total (8 frequencies × 4 spectral modulations × 4 temporal modulations, see Santoro et al. (2014) for details). Second, we employed this modulation-based representation to perform multivoxel decoding (as described by Eqs. (8–10)) in the identified PRs (and corresponding $\overline{\mathrm{PR}}$s). As for the *Weighted Pitch* model, these analyses resulted in a Pearson's correlation between reconstructed and original features (r_s), a normalized rank (m_s) per each sound in the test set and an average identification accuracy score per subject. Third, to examine whether the pitch model contributes relevant decoding information in addition to the *Modulation* model, we calculated a combined *Modulation-Pitch* identification score by averaging the correlation coefficients obtained separately for the *Weighted Pitch* and *Modulation* models and re-computing the normalized rank from the averaged correlation for each sound (Eq. (10)). Fisher transform/inverse transforms were applied before/after the averaging.

Finally, we compared the identification accuracy obtained with the combined *Modulation-Pitch* decoder with that obtained with the *Weighted Pitch* and the *Modulation* decoders by performing group-level (one-tailed) paired t-tests on the Fisher transformed values.

Results

Voxel-based prediction accuracy and model comparison

Fig. 2 shows the group maps of prediction accuracy obtained for all considered sound representation models. The *Weighted Pitch* model, which represented the pitch of natural sounds as a "weighted" combination of pitch height and salience, showed the highest prediction accuracy for cortical regions located bilaterally along HS, medial to HG, in lateral HG and adjacent regions in middle STG and in posterior STG (Fig. 2B). A similar distribution of accuracy values was observed for the *Pitch* model (Fig. 2C), where pitch was instead only modeled by the F0 contour. The *Tonotopy* model, which considered each sound's spectral energy, showed the most predictive power in voxels surrounding HG medially (in the first transverse sulcus [FTS]) and posteriorly (along HS) in both the hemispheres (Fig. 2D). Prediction accuracy of the *Timbral Brightness* model, which represented sounds by the height of the spectral centroid, followed the same arrangement as for the *Tonotopy* model but with lower overall values (Fig. 2E). The group maps of the prediction accuracy obtained for the *Saliency* model (Fig. 2F) showed lower overall values with respect to the *Pitch* and to the *Weighted Pitch* models. The prediction accuracy of the *Saliency-Pitch* model (Fig. 2G), instead, showed an arrangement similar to that of the *Pitch* model.

Fig. 3 shows the group-level statistical non-parametric maps comparing the *Weighted Pitch* to the competing models. The accuracies of the two pitch models did not differ significantly in the left hemisphere. In contrast, in the right hemisphere the *Weighted Pitch* model performed significantly better than the *Pitch* model in middle STG (at the lateral adjacency of HG/HS) and posterior STG (Fig. 3A). When compared to both the *Tonotopy* and the *Timbral Brightness* models, in both the left and right hemisphere, the *Weighted Pitch* model yielded significantly higher prediction accuracy on lateral HG and on adjacent STG regions, which also extended more posteriorly (Fig. 3B-C respectively). Furthermore, the *Weighted Pitch* model outperformed the *Saliency* model in middle and posterior STG bilaterally and in the surrounding area of HG of the left hemisphere (Fig. 3D) and the *Saliency-Pitch* model in right middle STG (at the lateral adjacency of HG/HS) and right posterior STG (Fig. 3E). These latter results were similar to those obtained in the comparison to the *Pitch* model.

Characterization of the Pitch ROI

Following these voxel-based comparisons, we performed additional analyses to test the hypothesis that the perceived pitch is encoded

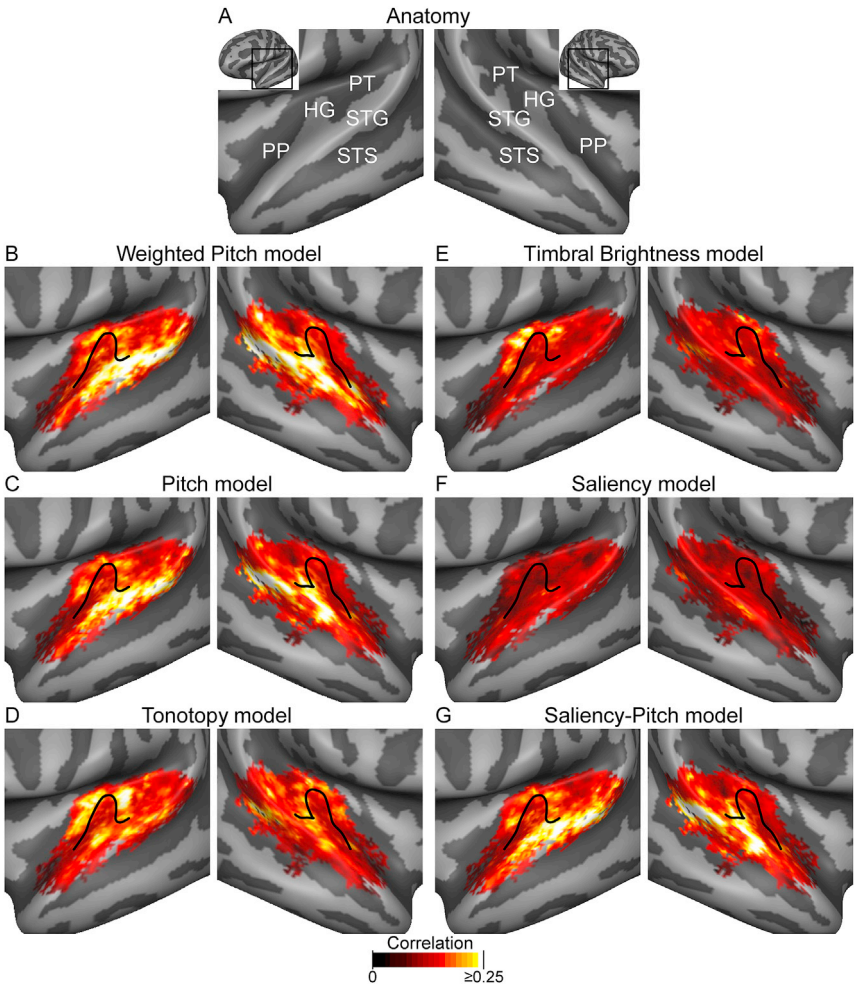

Fig. 2. Group maps of voxels' prediction accuracy projected on the inflated reconstruction of the group auditory cortex. (A) Surface reconstruction of the group auditory cortex. The black square in the insets illustrates which part of the complete cortical meshes is displayed (B–G). The prediction accuracy is quantified as the median value across subjects of the voxel-wise Pearson's correlation coefficient between measured and predicted fMRI responses to the testing sounds. Red [white] colors indicate low [high] accuracy correlation values. Black lines denote the HG.

preferentially within a specific region of the auditory cortex. For each single subject we defined a *Pitch ROI* (PR) using an automated procedure based on training data alone (see Materials and Methods). The location of the PR revealed a high consistency across subjects on lateral HG and middle STG bilaterally (Fig. 4A, see Fig. S1 for results at single subject level). These regions were consistent with the locations having the highest prediction accuracy of the fMRI activity to the testing sounds in the group analysis (Fig. 4B).

Fig. 4C shows the overlap of the PR with group tonotopy maps in the cortex-based realigned space. These maps were obtained as by color-coding the median value across subjects of voxels' characteristic frequency (CF), as estimated with the *Tonotopy* model (Moerel et al., 2012). CF maps showed a typical pattern with multiple low-high frequency gradients covering HG and surrounding STG (see Moerel et al. (2014) for a detailed description). The PR mostly matched the region with preference for low frequencies occupying the lateral Heschl's gyrus and adjacent STG.

When considering a multiresolution sound representation, the energy distribution over high spectral scales carries information about the pitch of a sound (see Wang and Shamma (1995) and Discussion). It is therefore interesting to evaluate the relation of the PR not only to tonotopic maps but also to maps of characteristic spectral modulations (CSM), as described in Santoro et al. (2014). Fig. 4D shows the overlap of the PR with the group CSM maps in the cortex-based realigned space. These maps were obtained by color-coding the median value across subjects of voxels' characteristic scale, as estimated with the *Modulation* model (Santoro et al., 2014). In accordance with previous results, CSM maps presented a preference for fast spectral scales in regions along HG and in anterior regions (Santoro et al., 2014; Schönwiesner and Zatorre, 2009). Interestingly, in both hemispheres, only a small portion of the PR included voxels preferring the fast scales (above 2.5 cyc/oct, purple colors). The remaining part of the PR corresponded to the area tuned to lower spectral modulation values, suggesting that the PR encodes a distinct representation of pitch (see Discussion).

Multivoxel decoding and combination with the modulation model

For both the identified PR and $\overline{\text{PR}}$, we quantified the capability of the *Weighted Pitch* model to correctly decode the perceived pitch of sounds from the multi-voxel patterns of brain activity by statistical assessment of the sound identification accuracy (see Materials and Methods). Accuracy was significantly above chance in both the ROIs (0.5, $p = 0.002$, two-sided signed rank test), but pitch identification was significantly more accurate within the PR (PR: mean [SEM] = 0.65 [0.021]; $\overline{\text{PR}}$: mean [SEM] = 0.59 [0.014]; $p = 0.006$, paired *t*-test; Fig. 5, see Table S1 for single subject results).

For these regions, we performed the same identification analysis for the *Modulation* and *Modulation-Pitch* decoders (see Materials and Methods). As expected (Santoro et al., 2014, 2017), in both the ROIs the identification accuracy for the *Modulation* decoder was significantly

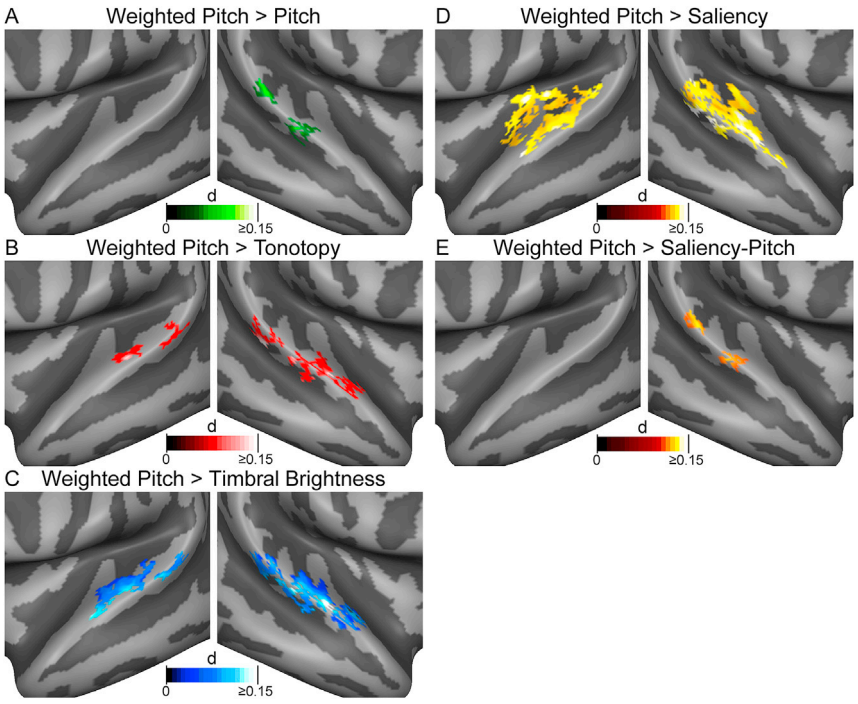

Fig. 3. Voxel-based model comparison. The contrast maps show the regions where the *Weighted Pitch* model significantly outperformed the competing models ($p < 0.05$; corrected for multiple comparisons using a cluster size correction). The color-code represents the value of the test statistic *d* defined for each voxel as the average of the difference between the prediction accuracy of the two corresponding models.

higher than chance (PR: mean [SEM] = 0.72 [0.020]; $\overline{PR}$: mean [SEM] = 0.69 [0.015], $p = 0.002$, two-sided signed rank test). In both the ROIs, the identification accuracy for the *Modulation* decoder was also significantly higher than the accuracy obtained with the *Weighted Pitch* decoder (PR: $p = 0.001$; $\overline{PR}$: $p = 4{\cdot}10^{-4}$; Fig. 5, see Table S2 for single subject results).

The combined *Modulation-Pitch* decoder provided highly significant identification accuracies within both the ROIs (PR: mean [SEM] = 0.74 [0.021]; $\overline{PR}$: mean [SEM] = 0.69 [0.015]; $p = 0.002$; Fig. 5, see Table S3 for single subject results). Importantly, in the PR the accuracy for the *Modulation-Pitch* decoder was significantly higher than the accuracy of both the *Weighted Pitch* decoder ($p = 1.2{\cdot}10^{-4}$) and the *Modulation* decoder ($p = 0.01$) (Fig. 5, left). Conversely, in the $\overline{PR}$, the accuracy for the *Modulation-Pitch* decoder was significantly higher than the accuracy of the *Weighted Pitch* decoder ($p = 1.2{\cdot}10^{-4}$) but not of the *Modulation* decoder ($p = 0.09$) (Fig. 5, right).

Discussion

In the present study we combined fMRI encoding/decoding with a computational algorithm of pitch extraction to investigate the representation of pitch of natural sounds in the human auditory cortex.

Our results showed that a model representing perceived pitch as a "*weighted*" combination of height and salience predicts the fMRI activity in distinct portions of the auditory cortex comparatively better than other perceptual and acoustic models. In particular, we found this effect to be most consistent across subjects in regions located in lateral HG and adjacent middle-posterior STG (*Pitch ROI*). This finding is in agreement with several previous fMRI studies that reported selective responses to pitch-evoking sounds in similar cortical locations (e.g. Griffiths and Hall, 2012; Patterson et al., 2002; Penagos et al., 2004). The agreement between previous and our fMRI findings is remarkable as the approaches differ in many respects. First, most of the studies so far entailed sets of synthetic stimuli with homogenous acoustic properties (e.g. IRN, harmonic complexes). Our stimuli, instead, consisted of a wide variety of real-life sounds that largely differed among each other both in terms of their acoustic (spectral and temporal) and perceptual properties. In this respect, the convergence of results obtained with simple artificial and complex real-life sounds suggest an overlap between the involved neural mechanisms. Second, in most previous studies, the localization of auditory cortical regions selective for pitch processing has been based on the statistical subtraction of the activation levels measured for the pitch evoking stimuli with those measured for stimuli designed to control e.g. for the influence of the spectral (or temporal) composition of the sound (Griffiths and Hall, 2012; Oxenham, 2013). While useful in cases where it is reasonable to assume that experimental and control sounds differ along a single dimension, this type of discriminative analysis becomes problematic with more complex stimuli. In fact, it is difficult to design control stimuli that are matched to real-life sounds in terms of acoustic and perceptual properties. With model-based fMRI such control stimuli are not required as the inference on (the localization of) pitch processing is based on the statistical assessment and comparison of alternative models in explaining/predicting fMRI responses. Note that the choice of the models to compare in fMRI encoding is as relevant as the choice of control stimuli in typical subtraction designs. In the present study, the pitch model based on the *weighted* combination of saliency and height (*Weighted Pitch* model) outperformed a model reflecting the spectral content of the sound (*Tonotopy* model). This suggests that the activity of neuronal populations in lateral HG and middle-posterior STG reflects pitch extraction and pitch representation in addition to the acoustic energy of the sound. The low performance of the *Timbral Brightness* model, instead, might depend on the fact that cortical responses may encode measures of temporal variability (e.g. interquartile range dissimilarity) rather than long-term statistics of the spectral centroid (Giordano et al., 2013).

Importantly, the *Weighted Pitch* model outperformed two separate models based on pitch height and salience alone, thus supporting the hypothesis that the auditory cortical responses in the identified regions reflect also the strength of pitch perception (Penagos et al., 2004). Previous studies indicated the involvement of regions of the right hemispheric auditory cortex in processing sound harmonicity (Giordano et al., 2013; Leaver and Rauschecker, 2010; Lewis et al., 2009). Accordingly,

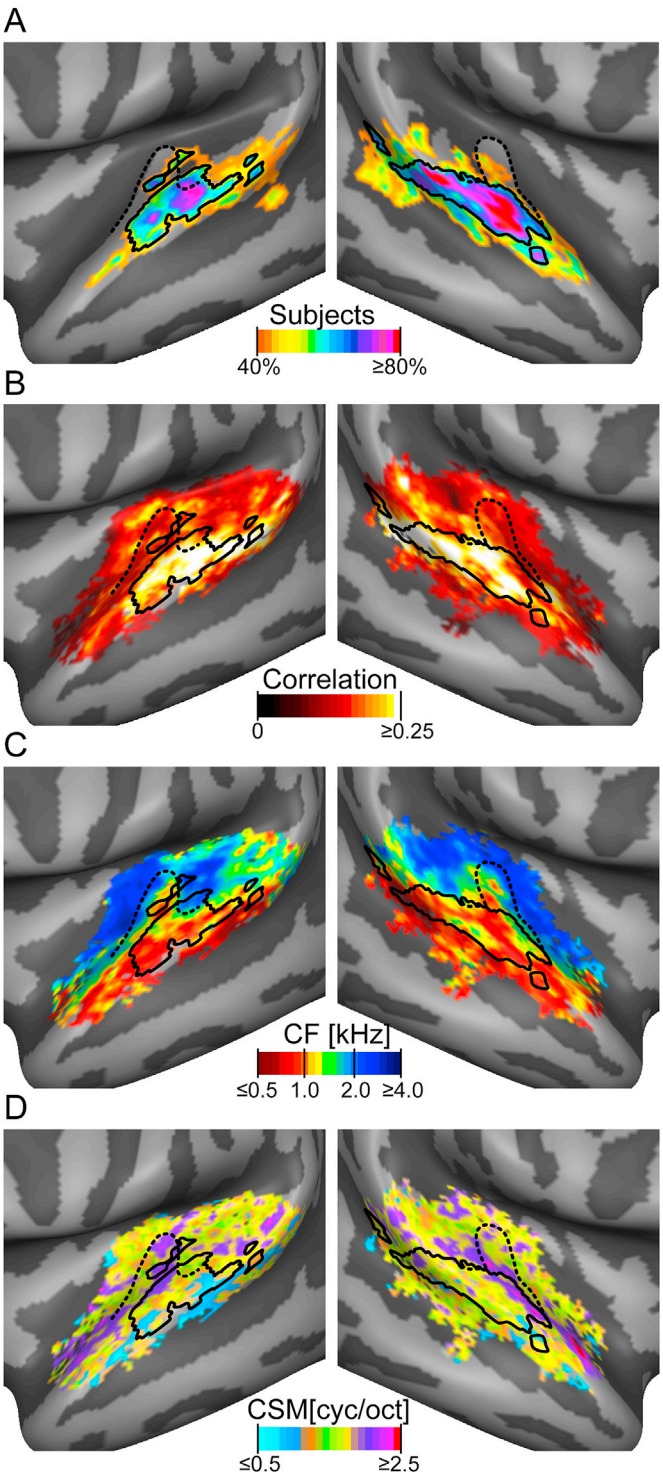

Fig. 4. (A) Consistency across subjects of the *Pitch ROI*. Orange [red] indicates an overlap of 40% [80%] across the subjects, respectively. (B) Overlap of the *Pitch ROI* with prediction accuracy group maps estimated for the *Weighted Pitch* model (i.e., the map displayed in Fig. 2B). (C,D) Pitch ROI superimposed to the characteristic frequency (CF) and spectral modulation (CSM) maps respectively. Panels (B–D) show only voxels active in at least 8 out of the 10 subjects. Black solid lines delineate the *Pitch ROI* corresponding to more than 60% overlap across subjects. Black dotted lines denote HG.

our *Saliency* model performed best on the STG of the right hemisphere. However, our findings support the relevance of pitch height in addition to salience in the encoding of natural sounds by the auditory cortex. In particular, in the highlighted auditory cortical regions, a modulatory (multiplicative) combination of salience and height in the *Weighted Pitch*

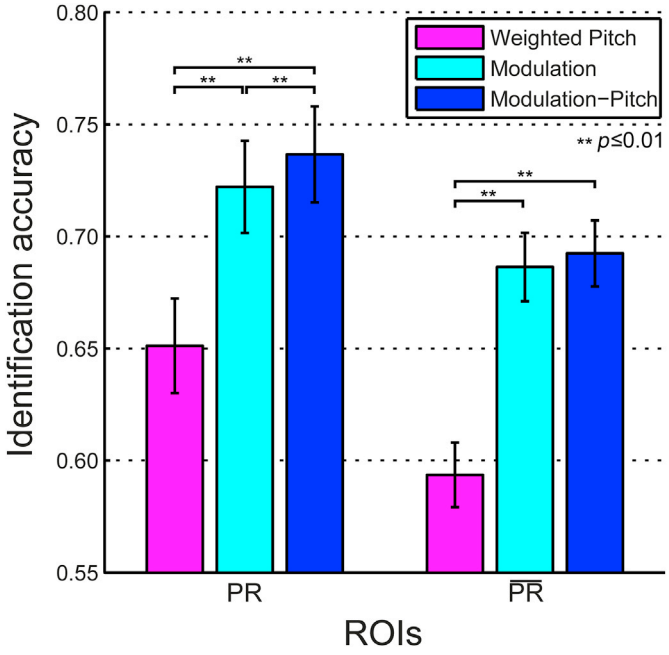

Fig. 5. Identification accuracy (mean ± SEM) obtained with the *Weighted Pitch, Modulation* and *Modulation-Pitch* decoders within the *Pitch ROI* (PR) and for the *Complementary Pitch ROI* ($\overline{\text{PR}}$). Horizontal lines indicate the significance of the pairwise comparisons.

model outperformed the simpler (additive) conjunction of height and salience information (*Saliency-Pitch* model) and the salience information alone (*Salience* model). We suggest that the *Weighted Pitch* model reflects a representation that more closely reflect the perception of the pitch of natural sounds.

Previous studies in the marmoset monkey reported that the largest number of "pitch-sensitive" neurons were located in a low-frequency region between A1 and lateral belt (Bendor and Wang, 2005). To examine the relation between the identified *Pitch ROI* and the auditory cortical tonotopic maps, we used the tonotopy model to derive topographic maps of voxels' characteristic frequency (CF) (Fig. 4C; see also Moerel et al. (2012)). The resulting maps followed the *tonotopic* organization of the human auditory cortex described in preceding imaging studies using fMRI (Da Costa et al., 2011; Formisano et al., 2003; Moerel et al., 2012; Saenz and Langers, 2014). Consistent with the findings in the marmoset monkey, the *Pitch ROI* overlapped substantially with the low frequency regions located in antero-lateral HG, but also extended into higher frequency clusters in middle/posterior STG.

Of the many existing algorithms of pitch extraction (e.g. de Cheveigné, 2005; Rabiner et al., 1976) we selected the YIN algorithm because it provided robust estimates of fundamental period (frequency) and harmonicity (aperiodicity) not only for speech sounds but also for higher pitched sounds of other categories. Whereas YIN is based on a temporal model of pitch (autocorrelation), our data and analyses do not allow making conclusions on whether pitch is extracted along the auditory system through temporal or spectral (spatial) mechanisms. In fact, only the output of the algorithm is used in the fMRI encoding/decoding analyses. Thus, using a different algorithm based on spectral analysis (Cohen et al., 1995; Shamma and Klein, 2000) would have affected our results only if the output representation would have been different. Similarly, we have formulated the fMRI encoding/decoding problem using a "spectral" representation of pitch, which was done in order to compare directly the pitch model to acoustic (tonotopic) models accounting for the sound spectral energy. Note that the fMRI responses could have been modeled equivalently (Eqs. (6) and (8)) in terms of a "temporal" representation of pitch. However, the nature of the fMRI signal does not allow resolving the temporal dynamics of the underlying

neuronal populations. Investigating the contribution of temporal mechanisms to the coding of pitch in complex sounds thus requires electrophysiological measurements (e.g. Bendor et al., 2012).

In our previous work we had examined the cortical processing of natural sounds by using either single-voxel encoding or multivariate decoding models. In particular, we adopted a single-voxel encoding approach to derive the profiles of voxels' sensitivity to physical acoustic features, such as frequency tuning curves (Moerel et al., 2012) and spectro-temporal modulation transfer functions (Santoro et al., 2014). Additionally, we employed a model-based multivariate decoding technique to further investigate how acoustic features (frequencies and modulations) are represented by patterns of activation within distinct auditory cortical regions (Santoro et al., 2017). Here we combined fMRI encoding and decoding as complementary techniques (Naselaris et al., 2011). Specifically, we first used the encoding approach to compare competing models of stimulus representation at single-voxel level. This comparison was done using group-level statistics based on non-parametric permutation testing and a cluster-based correction for multiple comparisons.

The assessment of single-voxel encoding results, however, relies on the spatial realignment of anatomical/functional data across subjects. Furthermore, the encoding model makes the assumption that the stimulus features that maximally contribute to a voxel response are also those encoded with greatest fidelity. But, higher responses might not necessarily mean better encoding and spatial response patterns may be informative of the pitch of complex sounds (Staeren et al., 2009). For these reasons, we complemented the single-voxel encoding with multivariate decoding, where data from individual voxels were jointly modeled. Whereas multivariate decoding is often limited to anatomically pre-defined ROIs, we implemented an automated procedure to define the *Pitch ROIs* at single-subject level. This enabled us to assess that spatial patterns of activation in the lateral HG and adjacent STG regions are indeed more informative of pitch height and salience compared to the complementary remainder of auditory cortex.

Additionally, our multivariate analyses showed that a decoder combining pitch and spectro-temporal modulation information is slightly but significantly more accurate than a decoder based on spectro-temporal modulation alone. This is consistent with current models of auditory scene analysis hypothesizing that the auditory system uses pitch in parallel to the multi-resolution representation for parsing the auditory objects of complex scenes (Elhilali and Shamma, 2008; Shamma et al., 2011). Note that within the modulation-based representation, the energy distribution over the spectral scales carries information about the pitch of a sound (Wang and Shamma, 1995). Harmonic sound components generate logarithmically spaced energy peaks in the spectral modulation scale-frequency plane, especially in the high-scale region (Zotkin et al., 2005). However, in such representation pitch is encoded only implicitly and obtaining an explicit pitch representation requires additional calculations. For example, an estimate of pitch can be obtained from the slope of the straight line that connects the scale peaks (Wang and Shamma, 1995) or using an algorithm based on spectral analysis (Shamma and Klein, 2000). Our results showing that the Pitch ROI only marginally overlaps with the regions preferring high spectral scales (Fig. 4D) is consistent with the hypothesis that the Pitch ROI encodes an explicit pitch representation, which may be the result of such calculations. In conclusion, our model based analysis of fMRI responses demonstrates that auditory cortical regions that have been implicated in the analysis of the pitch of simple synthetic sounds also represent the pitch of complex real life sounds. Furthermore, our results suggest that these representations do not only encode perceived pitch height but also perceived pitch saliency.

Acknowledgments

This work was supported by Maastricht University, the Netherlands Organisation for Scientific Research (VICI grant 453-12-002 to E.F, VIDI grant 864-13-012 to F.D.M., VENI grant 451-15-012 to M.M) and the Dutch Province of Limburg.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2017.11.020.

References

Barker, D., Plack, C.J., Hall, D.A., 2012. Reexamining the evidence for a pitch-sensitive region: a human fMRI study using iterated ripple noise. Cereb. Cortex 22, 745–753.
Bendor, D., 2012. Does a pitch center exist in auditory cortex? J. Neurophysiol. 107, 743–746.
Bendor, D., Osmanski, M.S., Wang, X., 2012. Dual-pitch processing mechanisms in primate auditory cortex. J. Neurosci. 32, 16149–16161.
Bendor, D., Wang, X., 2005. The neuronal representation of pitch in primate auditory cortex. Nature 436, 1161–1165.
Bendor, D., Wang, X., 2006. Cortical representations of pitch in monkeys and humans. Curr. Opin. Neurobiol. 16, 391–399.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc.
Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press.
Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887–906.
Cohen, M.A., Grossberg, S., Wyse, L.L., 1995. A spectral network model of pitch perception. J. Acoust. Soc. Am. 98, 862–879.
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.
de Cheveigné, A., 2005. Pitch perception models. In: Plack, C.J., Fay, R.R., Oxenham, A.J., Popper, A.N. (Eds.), Pitch: Neural Coding and Perception. Springer New York, pp. 169–233.
de Cheveigné, A., 2010. Pitch perception. In: Plack, C.J. (Ed.), The Oxford Handbook of Auditory Science: Hearing. Oxford University Press, pp. 71–104.
de Cheveigné, A., Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. J. Acoust. Soc. Am. 111, 1917–1930.
De Martino, F., Moerel, M., van de Moortele, P.-F., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E., 2013. Spatial organization of frequency preference and selectivity in the human inferior colliculus. Nat. Commun. 4, 1386.
Elhilali, M., Shamma, S.A., 2008. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J. Acoust. Soc. Am. 124, 3751–3771.
Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn. Reson Med. 33, 636–647.
Formisano, E., Kim, D.-S., Di Salle, F., van de Moortele, P.-F., Ugurbil, K., Goebel, R., 2003. Mirror-symmetric tonotopic maps in human primary auditory cortex. Neuron 40, 859–869.
Garcia, D., Hall, D.A., Plack, C.J., 2010. The effect of stimulus context on pitch representations in the human auditory cortex. NeuroImage 51, 808–816.
Giordano, B.L., McAdams, S., Zatorre, R.J., Kriegeskorte, N., Belin, P., 2013. Abstract encoding of auditory objects in cortical activity patterns. Cereb. Cortex 23, 2025–2037.
Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum. Brain Mapp. 27, 392–401.
Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223.
Griffiths, T.D., Buchel, C., Frackowiak, R.S.J., Patterson, R.D., 1998. Analysis of temporal structure in sound by the human brain. Nat. Neurosci. 1, 422–427.
Griffiths, T.D., Hall, D.A., 2012. Mapping pitch representation in neural ensembles with fMRI. J. Neurosci. 32, 13343–13347.
Griffiths, T.D., Uppenkamp, S., Johnsrude, I., Josephs, O., Patterson, R.D., 2001. Encoding of the temporal regularity of sound in the human brainstem. Nat. Neurosci. 4, 633–637.
Hall, D.A., Plack, C.J., 2007. The human 'pitch center' responds differently to iterated noise and Huggins pitch. NeuroReport 18, 323–327.
Hall, D.A., Plack, C.J., 2009. Pitch processing sites in the human auditory brain. Cereb. Cortex 19, 576–585.
Hoerl, A.E., Kennard, R.W., 1970. ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Houtsma, A.J.M., 1997. Pitch and timbre: definition, meaning and use. J. New Music Res. 26, 104–115.
Kay, K.N., David, S.V., Prenger, R.J., Hansen, K.A., Gallant, J.L., 2008a. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum. Brain Mapp. 29, 142–156.
Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008b. Identifying natural images from human brain activity. Nature 452, 352–355.
Kendall, R.A., Carterette, E.C., Hajda, J.M., 1999. Perceptual and acoustical features of natural and synthetic orchestral instrument tones. Music Percept. Interdiscip. J. 16, 327–363.

Leaver, A.M., Rauschecker, J.P., 2010. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. 30, 7604–7612.
Lewis, J.W., Talkington, W.J., Walker, N.A., Spirou, G.A., Jajosky, A., Frum, C., 2009. Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. J. Neurosci. 29, 2283–2296.
Marozeau, J., de Cheveigne, A., McAdams, S., Winsberg, S., 2003. The dependency of timbre on fundamental frequency. J. Acoust. Soc. Am. 114, 2946–2957.
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J., 1995. Perceptual scaling of synthesized musical timbres: common dimensions, specificities, and latent subject classes. Psychol. Res. 58, 177–192.
Meddis, R., Hewitt, M.J., 1991. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: pitch identification. J. Acoust. Soc. Am. 89, 2866–2882.
Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.A., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915–929.
Moerel, M., De Martino, F., Formisano, E., 2012. Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity. J. Neurosci. 32, 14205–14216.
Moerel, M., De Martino, F., Formisano, E., 2014. An anatomical and functional topography of human auditory cortical areas. Front. Neurosci. 8, 225.
Moerel, M., De Martino, F., Santoro, R., Ugurbil, K., Goebel, R., Yacoub, E., Formisano, E., 2013. Processing of natural sounds: characterization of multipeak spectral tuning in human auditory cortex. J. Neurosci. 33, 11888–11898.
Moore, B.C.J., 1995. Hearing. Academic Press.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410.
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.
Oxenham, A.J., 2013. Revisiting place and temporal theories of pitch. Acoust. Sci. Technol. 34, 388–396.
Oxenham, A.J., Bernstein, J.G.W., Penagos, H., 2004. Correct tonotopic representation is necessary for complex pitch perception. Proc. Natl. Acad. Sci. U. S. A. 101, 1421–1425.
Patterson, R., Handel, S., Yost, W.A., Datta, A.J., 1996. The relative strength of the tone and noise components in iterated rippled noise. J. Acoust. Soc. Am. 100, 3286–3294.
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.
Plack, C.J., Oxenham, A.J., Fay, R.R., Popper, A.N., 2005. Pitch: Neural Coding and Perception. Springer New York.
Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C., 1976. A comparative performance study of several pitch detection algorithms. IEEE Trans. Acoust. Speech, Signal Process. 24, 399–418.
Saenz, M., Langers, D.R.M., 2014. Tonotopic mapping of human auditory cortex. Hear. Res. 307, 42–52.
Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. 10, e1003412.
Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., Formisano, E., 2017. Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc. Natl. Acad. Sci. 114, 4799–4804.
Schönwiesner, M., Zatorre, R.J., 2009. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. 106, 14611–14616.
Shamma, S., Klein, D., 2000. The case of the missing pitch templates: how harmonic templates emerge in the early auditory system. J. Acoust. Soc. Am. 107, 2631–2644.
Shamma, S.A., 2004. Topographic organization is essential for pitch perception. Proc. Natl. Acad. Sci. U. S. A. 101, 1114–1115.
Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci. 34, 114–123.
Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. 19, 498–502.
Wang, K., Shamma, S.A., 1995. Spectral shape analysis in the central auditory system. IEEE Trans. Speech Audio Process. 3, 382–395.
Yost, W.A., Patterson, R., Sheft, S., 1996. A time domain description for the pitch strength of iterated rippled noise. J. Acoust. Soc. Am. 99, 1066–1078.
Zotkin, D.N., Chi, T., Shamma, S.A., Duraiswami, R., 2005. Neuromimetic sound representation for percept detection and manipulation. EURASIP J. Adv. Signal Process. 2005, 486137.