Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns

Roberta Santoro[a,b,c,1], Michelle Moerel[a,b,d,e], Federico De Martino[a,b], Giancarlo Valente[a,b], Kamil Ugurbil[d], Essa Yacoub[d], and Elia Formisano[a,b,e,1,2]

[a]Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, 6200 MD Maastricht, The Netherlands; [b]Maastricht Brain Imaging Center, 6200 MD Maastricht, The Netherlands; [c]Brain and Language Laboratory, Department of Clinical Neuroscience, University Medical School, University of Geneva, CH-1211 Geneva, Switzerland; [d]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455; and [e]Maastricht Centre for Systems Biology, Maastricht University, 6200 MD Maastricht, The Netherlands

Edited by David Poeppel, New York University, New York, NY, and accepted by Editorial Board Member Michael S. Gazzaniga March 24, 2017 (received for review October 25, 2016)

Ethological views of brain functioning suggest that sound representations and computations in the auditory neural system are optimized finely to process and discriminate behaviorally relevant acoustic features and sounds (e.g., spectrotemporal modulations in the songs of zebra finches). Here, we show that modeling of neural sound representations in terms of frequency-specific spectrotemporal modulations enables accurate and specific reconstruction of real-life sounds from high-resolution functional magnetic resonance imaging (fMRI) response patterns in the human auditory cortex. Region-based analyses indicated that response patterns in separate portions of the auditory cortex are informative of distinctive sets of spectrotemporal modulations. Most relevantly, results revealed that in early auditory regions, and progressively more in surrounding regions, temporal modulations in a range relevant for speech analysis (∼2–4 Hz) were reconstructed more faithfully than other temporal modulations. In early auditory regions, this effect was frequency-dependent and only present for lower frequencies (<∼2 kHz), whereas for higher frequencies, reconstruction accuracy was higher for faster temporal modulations. Further analyses suggested that auditory cortical processing optimized for the fine-grained discrimination of speech and vocal sounds underlies this enhanced reconstruction accuracy. In sum, the present study introduces an approach to embed models of neural sound representations in the analysis of fMRI response patterns. Furthermore, it reveals that, in the human brain, even general purpose and fundamental neural processing mechanisms are shaped by the physical features of real-world stimuli that are most relevant for behavior (i.e., speech, voice).

auditory cortex | functional MRI | natural sounds | model-based decoding | spectrotemporal modulations

Many natural and man-made sources in our environment produce acoustic waveforms (sounds) consisting of complex mixtures of multiple frequencies (Fig. 1*A*). At the cochlea, these waveforms are decomposed into frequency-specific temporal patterns of neural signals, typically described with auditory spectrograms (Fig. 1*B*). How complex sounds are further transformed and analyzed along the auditory neural pathway and in the cortex remains uncertain. Ethological considerations have led to the hypothesis that brain processing of sounds is optimized for spectrotemporal modulations, which are characteristically present in ecologically relevant sounds (1), such as in animal vocalizations [e.g., zebra finch (2), macaque monkeys (3)] and speech (2, 4–7). Modulations are regular variations of energy in time, in frequency, or in time and frequency simultaneously. Typically, in natural sounds, modulations dynamically change over time. The contribution of modulations to sound spectrograms can be made explicit using 4D (time-varying; Movies S1–S3, *Left*) or 3D (time-averaged; Fig. 1*C*) representations. Such representations highlight, for example, that the energy of human vocal sounds is mostly concentrated at lower frequencies and at lower temporal modulation rates (Fig. 1*C*, *Left*), whereas the whinny of a horse contains energy at higher frequencies and at higher temporal modulations (Fig. 1*C*, *Right*).

Electrophysiological investigations in several animal species have reported single neurons tuned to specific spectrotemporal modulations at various stages of the auditory pathway [e.g., inferior colliculus (8), auditory thalamus (9)] and in the primary auditory cortex (10, 11). Intracranial electrocorticography (ECoG) recordings (12, 13) as well as noninvasive functional neuroimaging studies (14, 15) suggest that similar mechanisms are also in place in the human auditory cortex.

In the present study, we tested the hypothesis that the human auditory cortex entails modulation-based sound representations by combining real-life sound stimuli, high spatial resolution (7 Tesla) functional magnetic resonance imaging (fMRI), and the analytical approach of model-based decoding (16–20). Unlike the more common classification-based decoding, which only allows discriminating between a small set of stimulus categories, this approach embeds a representational model of the stimuli in terms of elementary features, thereby enabling the identification of individual arbitrary stimuli from brain response patterns (18, 19).

Specifically, we modeled the cortical processing of real-life sounds as the combined output of frequency-localized neural filters tuned to specific combinations of spectral and temporal modulations (14, 21). We then used this sound representation model in two sets of fMRI data analysis. In a first set of analyses,

Significance

The sounds we encounter in everyday life (e.g. speech, voices, animal cries, wind, rain) are complex and various. How the human brain analyses their acoustics remains largely unknown. This research shows that mathematical modelling in combination with high spatial resolution functional magnetic resonance imaging enables reverse engineering of the human brain computations underlying real-life listening. Importantly, the research reveals that even general auditory processing mechanisms in the human brain are optimized for fine-grained analysis of the most behaviorally relevant sounds (i.e., speech, voices). Most likely, this observation reflects the evolutionary feat that, for humans, discriminating between speech sounds is more crucial than distinguishing, for example, between barking dogs.

Author contributions: R.S., M.M., F.D.M., and E.F. designed research; R.S., M.M., F.D.M., and E.Y. performed research; R.S., F.D.M., G.V., and E.F. contributed new reagents/analytic tools; R.S., M.M., F.D.M., and E.F. analyzed data; and R.S., M.M., F.D.M., G.V., K.U., E.Y., and E.F. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. D.P. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The fMRI data and stimuli have been deposited in Dryad (dx.doi.org/10.5061/dryad.np4hs).

[1]R.S. and E.F. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: e.formisano@maastrichtuniversity.nl.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617622114/-/DCSupplemental.

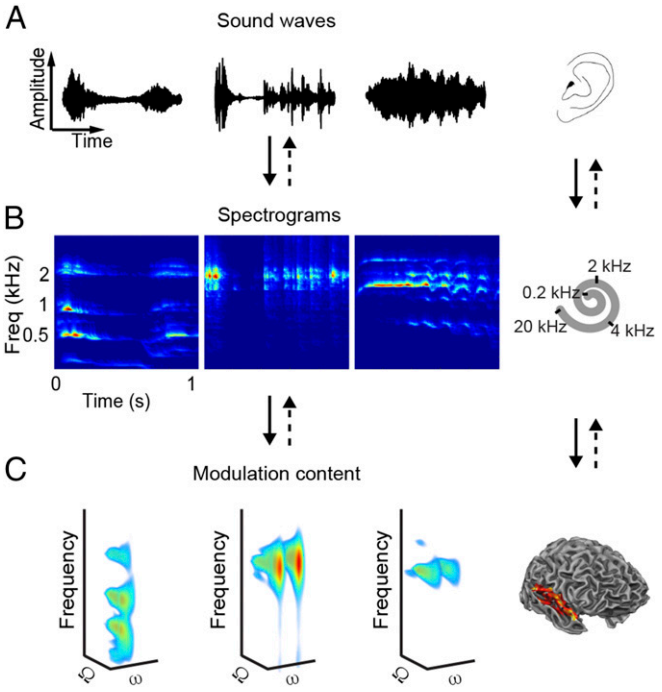

Fig. 1. Schematic of reconstruction procedure. Sounds enter a listener's ear as waveforms of acoustic energy (*A*) and are converted into spectrograms (*B*) at the cochlea. Freq, frequency. (*C*) Sound representations in the auditory cortex are modeled as 4D functions of frequency, spectral modulation (Ω), temporal modulation (ω), and time. Note that *C* shows the time-averaged representations of the sounds in *A* and *B* (time-resolved representations are illustrated in Movies S1–S3). Sounds are examples of human vocal (*Left*, "male vocalization"), tool (*Center*, "typewriter"), and animal (*Right*, "horse") sounds. The fMRI activation patterns are used to decode each feature (Ω, ω, frequency, time) of the 4D representation, which is then inverted (dashed arrows) to obtain the corresponding spectrogram and waveform.

conducted at the level of the whole auditory cortex, we trained a pattern-based decoder independently for each model feature (i.e., for each unique combination of frequency, spectral modulation, and temporal modulation). We then asked whether the combination of these feature-specific decoders would enable reconstruction of the acoustic content of hold-out sets of sounds from fMRI response patterns. Supporting the hypothesis embedded in our model, obtained reconstructions were significantly accurate and specific. Most surprisingly, despite the inherent loss of temporal information due to the sluggish hemodynamics and poor temporal sampling of the blood oxygen level-dependent (BOLD) response, fMRI-based reconstructions presented a temporal specificity of about 200 ms. In a second set of analyses, we considered the contribution of different auditory cortical regions separately and characterized each region by the accuracy of the fMRI-based reconstruction for each feature of the sound representation model, which we refer to as modulation transfer function (MTF). A detailed comparison of these regional MTFs revealed relevant insights into the processing of acoustic information in primary and nonprimary auditory cortical regions. Most interestingly, our results suggested that even in primary regions, sound representations and processing are optimized for the fine-grained discrimination of human speech (and vocal) sounds.

Results

Sound Reconstruction from fMRI Activity Patterns. We recorded 7-T fMRI responses from the auditory cortex while subjects listened to a large set of real-life sounds, including speech and vocal samples, music pieces, animal cries, scenes from nature, and tool sounds [experiment 1: $n_1 = 5$ (14, 22), experiment 2: $n_2 = 5$]. As a first step of the model-based decoding analysis, we calculated the time-varying spectrotemporal modulation content of all our stimuli (e.g., Movies S1–S3, *Left*). Then, per subject, we estimated a linear decoder for each feature of this modulation representation. This estimation was done using a subset of sounds and corresponding fMRI responses (training) and resulted in a map of voxels' contributions for each feature (C_i in Eq. **S1**). We then tested whether the combination of estimated feature-specific decoders could be used to reconstruct the time-varying spectrotemporal modulations of novel ("test") sounds based on the measured fMRI responses to those testing sounds. We refer to this operation as modulation-based reconstruction (*SI Materials and Methods*). We verified the quality of obtained sound reconstructions by means of several statistical analyses.

First, we assessed the accuracy of the sounds' reconstructed modulation content using the coefficient of determination (R^2_{pred}) (Eq. **S3**). For each sound, R^2_{pred} is greater than 0 if the reconstructed modulation representation predicts the actual representation better than the mean of that sound. In both experiments, R^2_{pred} was significantly higher than 0 (Fig. S1*A*; experiment 1: median [interquartile range (IQR)] = 0.37 [0.36 0.47], $P < 0.05$; experiment 2: IQR = 0.44 [0.42 0.45], $P < 0.05$). (Unless differently indicated, statistical comparisons are based on random effects, group-level, one-tailed Wilcoxon signed-rank tests.) Conversely, a reconstruction model based on a time-frequency representation of the stimuli yielded poorer reconstruction accuracy, with distributions largely overlapping or only marginally shifted with respect to the null distribution [experiment 1: −0.05 [−0.06 0.01], not significant (n.s.); experiment 2: 0.02 [0.01 0.03], $P < 0.05$; Fig. S1*B*]. These results suggest that the modulation representation is crucial for the decoding of complex sounds from fMRI activity patterns.

Second, we assessed the specificity of the reconstructed modulation representations by examining to what extent the fMRI-based predictions enabled identifying a given sound among all testing sounds in the test set (*SI Materials and Methods*). Identification accuracy was significantly above chance (0.5) for both datasets on a group level (experiment 1: 0.78 [0.73 0.84], $P < 0.05$; experiment 2: 0.82 [0.79 0.83], $P < 0.05$) and on a single-subject level (for each subject: $P < 0.01$, one-tailed permutation test; Fig. 2*A* and *B*). Notably, in about 30% of cases, the identification scores were in the range of 0.9–1.0 (Fig. 2*B*), indicating that the fine-grained, within-category distinction between sounds contributed relevantly to the median score. Finally, because our modulation-based sound representations were based on a temporal subdivision of the sounds in 10 time windows, we further examined the temporal specificity of our fMRI-based reconstructions. We calculated separately for each time window (100 ms) the identification accuracy score using fMRI-based predictions corresponding to the same time window of the actual sound features (lag = 0; blue lines in Fig. S2) or to time windows at a distance of 1 (lag = 1; red lines in Fig. S2) or 2 (lag = 2; green lines in Fig. S2). Identification accuracy for lag = 0 did not differ significantly from the identification accuracy obtained for lag = 1 (experiment 1: n.s., experiment 2: n.s.), but it was significantly greater than the identification accuracy at lag = 2 (experiment 1: $P < 0.05$, experiment 2: $P < 0.0001$, random effects Friedman test with lag and subject as factors). This analysis suggests a temporal specificity of at least two time bins (i.e., 200 ms) for the obtained fMRI-based predictions.

To obtain an intuitive understanding of these results, we reconstructed spectrograms from the fMRI-derived modulation representations (Movies S1–S3, *Right*) and resynthesized the corresponding waveforms (*SI Materials and Methods*). As illustrated in Fig. 3 *A*–*C*, we could recover "temporally smooth" versions of the original spectrograms. In line with these results, resynthesized waveforms enabled the recognition of the original sound sources (Audio File S1, "bird" reconstruction) in some cases, but lacked the fine temporal details required, for example, for speech comprehension (Audio File S2, "speech"). A formal statistical analysis showed that recovery specificity for spectrograms was significantly

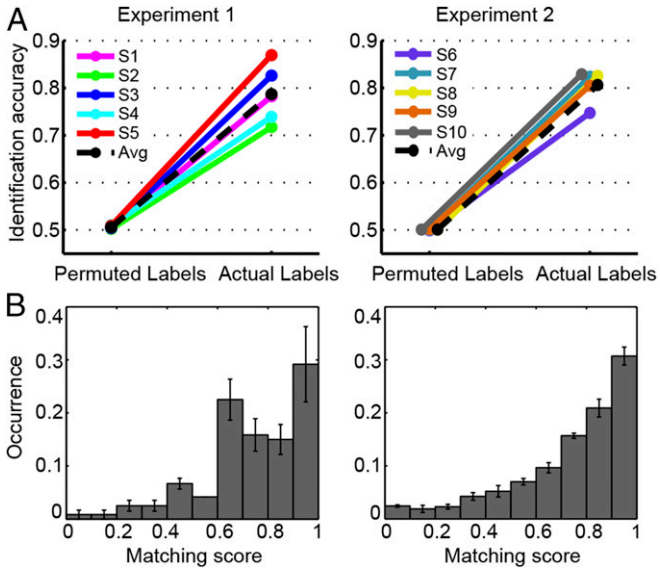

Fig. 2. Identification results. (*A*) Identification accuracy for individual participants. Each panel shows the accuracy obtained with correct labels and the accuracy derived by permuting the sound labels. (*B*) Average distribution of matching scores across subjects for the modulation-based reconstruction (mean ± SEM, $n = 5$, for the two experiments separately). These matching scores are used to calculate the identification accuracy. Avg, average.

above chance (0.5) (experiment 1: 0.65 [0.33 0.91], $P < 0.05$; experiment 2: 0.75 [0.43 0.90], $P < 0.05$).

Region of Interest Analysis of Spectral and Temporal Information Content. Having established that the modulation-based model enables sufficiently accurate and specific sound reconstructions, we investigated how the decoding of the spectrotemporal modulation content varies throughout auditory cortical regions. We compared reconstruction performance of six anatomical regions of interest (ROIs): Heschl's gyrus (HG), planum polare (PP), planum temporale (PT), anterior superior temporal gyrus (aSTG), middle STG (mSTG), and posterior STG (pSTG) (Fig. 4*A* and *SI Materials and Methods*). For each subject and ROI, we estimated the multivoxel decoders and quantified, per each feature, the reconstruction accuracy as Pearson's correlation coefficient (r) between predicted and actual feature values in all sounds of the test set. This procedure resulted in an MTF per ROI (Fig. 4*B*), with corresponding marginal frequency (f), spectral modulation (Ω), and temporal modulation (ω) profiles (Fig. 5). These MTFs were assessed statistically and thresholded ($P < 0.05$, corrected for multiple comparisons; *SI Materials and Methods*), and thereby provide an objective measure of what information about each feature of the model is available in the ROI's response patterns.

Results indicated that in the HG, PP, and PT, a broader range of acoustic features can be decoded compared with STG regions (Fig. 4*B*; detailed pairwise comparisons between ROIs are shown in Fig. S3). The reconstruction accuracy profile for frequency was highest at around 0.8 kHz for the HG, PT, and PP and at lower frequencies for the aSTG, mSTG, and pSTG (Fig. 5). For the HG, PT, and PP, reconstruction accuracy for frequencies above 2 kHz and below 0.5 kHz was significantly higher than in the frequency range between 0.5 and 2 kHz (Bonferroni-adjusted $P < 0.001$). Note that this behavior for frequency was not present in the stimuli (Fig. S4 *A* and *B*) and might be related to direct and indirect effects of the scanner noise (*Discussion* and Fig. S5). For the aSTG, mSTG, and pSTG, the reconstruction accuracy below 0.6 kHz was significantly greater than at higher frequencies (Bonferroni-adjusted $P < 0.001$). In all ROIs, the reconstruction accuracy profile for spectral modulations was highest for lower modulations (with a steeper slope at higher frequencies in the HG, PT, and PT; Fig. 5), with reconstruction accuracies at the lowest spectral modulations [0.5 cycles per octave (cyc/oct)] significantly higher than at four cyc/oct (Bonferroni-adjusted $P < 0.001$). Thus, there was not a preferred range of spectral modulations, because brain responses followed the spectral modulation content of the stimuli (Fig. S4). Conversely, in all regions, the temporal modulation profile was highest for a range centered at ~3 Hz (Fig. 5). Reconstruction accuracy at 3.1 Hz was significantly higher than at 1 Hz and 9.7 Hz for the aSTG, mSTG, and pSTG (Bonferroni-adjusted $P < 0.001$). Visual inspection of the MTFs (Fig. 4*B*) indicated that for the HG, PP, and PT, this effect was only observed in the frequency range below 2 kHz. Formal statistical testing confirmed this observation (Bonferroni-adjusted $P < 0.001$). Instead, at frequencies above 2 kHz, reconstruction accuracy of temporal modulations for HG, PP, and PT was higher at 30 Hz than at 1 Hz (Bonferroni-adjusted $P < 0.001$). This distinctive reconstruction accuracy profile for temporal modulations could not be explained by the overall acoustic properties of the stimuli (Fig. S4).

ROI-Based Analysis Without Speech and Vocal Stimuli. The low temporal modulation rates for which we have obtained the highest reconstruction accuracy are prominently present in speech and

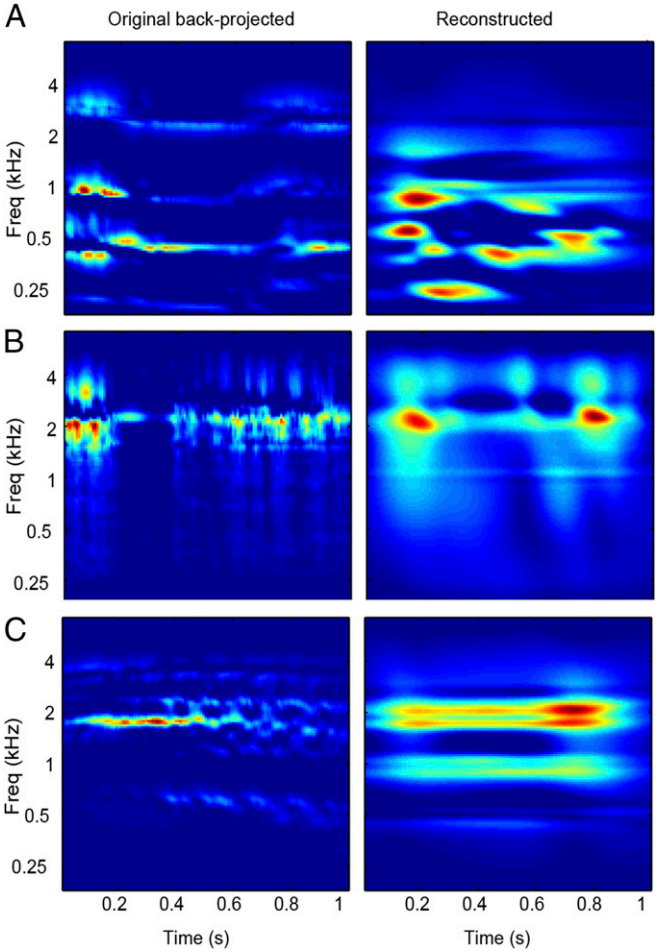

Fig. 3. Examples of reconstructed spectrograms. Reconstructed spectrograms for vocal (*A*), tool (*B*), and animal (*C*) sounds. The original spectrograms are depicted in Fig. 1. (*Left*) Reference spectrograms obtained by inverting the down-sampled (10 time bins) magnitude-only modulation representation of the original sounds (*SI Materials and Methods*). (*Right*) Sound spectrogram as reconstructed from the fMRI-based predictions (also Movies S1–S3).

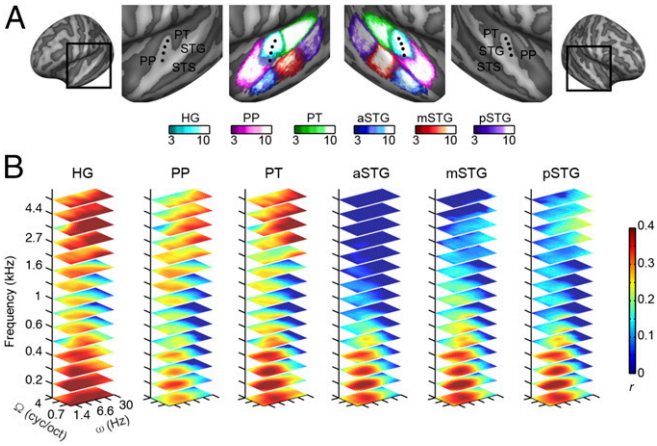

Fig. 4. MTFs of individual ROIs. (*A*) Inflated representation of the group cortical surface mesh. ROIs are shown with different colors scaled to indicate the overlap of defined regions across subjects. The black dots indicate the HG. (*B*) Each slice represents the MTF at a given frequency value (only 15 frequencies are shown). The color code indicates the group-averaged Pearson's *r* between reconstructed and original features. Features with a nonsignificant *r* are depicted in dark blue. The average MTF across hemispheres is shown. MTFs have been interpolated for display purposes. cyc/oct, cycles per octave.

vocal sounds, and are relevant for the analysis of syllabic information (23) and for speech intelligibility (6), for example. In many previous studies, processing of speech/voice (24–26) beyond the analysis of acoustic features has been related to stronger fMRI responses relative to other natural and control sounds, especially along the STG (and adjacent superior temporal sulcus). We thus performed a further analysis to control that the enhanced reconstruction accuracy of low temporal rates was not an indirect effect of decoding these global response differences. In this analysis, we removed all speech and vocal sounds from the stimulus set, retrained the decoders, and statistically reassessed the regional MTFs. Removal of speech and vocal sounds largely altered the relative contribution of low and high frequencies to the overall acoustic energy of the stimulus set used for training/testing the decoders, but affected temporal and spectral modulations less (Fig. S4). Fig. S6*A* shows, for all ROIs, the MTFs obtained with this reduced stimulus set and thresholded as in the previous analysis. The newly identified MTFs resembled the original ones very closely in early auditory regions (Pearson's *r* with the original MTFs, 3,600 features: HG = 0.90, PT = 0.83, PP = 0.89) and with larger deviations in STG regions [Pearson's *r*, 3,600 features: aSTG = 0.47, mSTG = 0.42, pSTG = 0.48]. In these latter regions, changes were most pronounced at the low frequencies. Importantly, in all ROIs, including early auditory as well as STG regions, the marginal profile for temporal modulation reconstruction accuracy remained unchanged (Fig. S6*B*), with a peak around 3 Hz (Pearson's *r*, over the 10 temporal modulations: HG = 0.99, PT = 0.99, PP = 0.98, aSTG = 0.97, mSTG = 0.97, pSTG = 0.95).

Discriminability Analysis. Electrophysiological recordings in zebra finches suggested that spectrotemporal population tuning of auditory neurons maximizes the acoustic distance between sounds, facilitating the animal's discrimination ability (27, 28). Under the assumption that the reconstructed accuracy of MTFs reflects the weighted distribution of neuronal populations tuned to the corresponding spectrotemporal modulation (*SI Discussion*), we tested the effects of this ROI-specific processing on sound discriminability. We calculated pairwise distances between sounds based on their original modulation representations as well as on the representations obtained by weighting the original representations based on the ROI-specific reconstruction accuracy of MTFs (Figs. S7 and S8). Comparison of these two sets of pairwise distances across the entire stimulus set exhibited a complex pattern (Fig. S8*A*). However, a clear pattern emerged when the analysis was restricted to speech sounds (Fig. S7*A*). In all ROIs, we found a significant linear relationship between the normalized distances of filtered and original speech sounds (slope of the regression line: HG = 1.09, PP = 1.03, PT = 1.16, aSTG = 1.40, mSTG = 1.38, pSTG = 1.38; Bonferroni-adjusted $P < 0.001$, two-tailed t test). Pairwise distances for all regions fell mainly above the diagonal (slope significantly higher than 1; Bonferroni-adjusted $P < 0.001$, two-tailed z test), indicating significant amplification of sound distances. The regression line was significantly steeper for the PT than for the HG and for the aSTG, mSTG, and pSTG than for the HG and PT (Bonferroni-adjusted $P < 0.001$, two-tailed z test). We did not observe similar effects for the other sound categories (Fig. S8 *B–E*), except for adults' vocal sounds (Fig. S7*B*).

Discussion

We applied an approach to embed models of neural sound representations in the analysis of fMRI response patterns, and thereby showed that it is feasible to reconstruct, with significant accuracy and specificity, the spectrotemporal modulation content of real-life sounds from fMRI signals. Successful decoding of the (time-averaged) spectral components of sounds could be expected based on the spatial organization of frequency in the auditory cortex (29, 30). Our current findings indicate that spectrotemporal sound modulations also map into distinct and reproducible spatial fMRI response patterns. This result is consistent with the hypothesis of a spatial representation of acoustic features besides frequency in primate (31) and human (14, 15, 32) auditory cortex.

The temporal specificity of obtained predictions indicated that not only the time-averaged modulation content of sounds but also modulation changes on the order of about 200 ms could be decoded from fMRI response patterns. Although consistent with recent reports of speech spectrogram reconstruction from ECoG recordings (12), this result is surprising, given the low temporal resolution of fMRI and the coarse temporal sampling [repetition time (TR) = 2.6 s] of brain responses. How is our result possible? We trained many (n = 153,600) multivariate decoders based on the same estimates of fMRI responses to the sounds. For each feature and for each time window, training resulted in a unique weighting of voxels' responses (i.e., the weights C_i). In other words, different sets of voxels were weighted relatively high or low for predictions corresponding to different features and time windows. This result suggests a mechanism by which spatial fMRI patterns are informed by temporal aspects of the sounds (also ref. 33). This

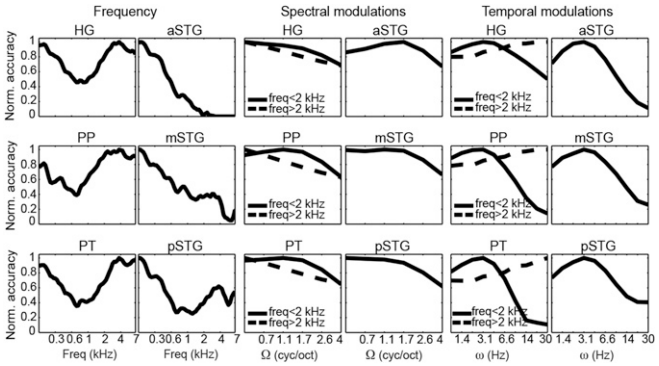

Fig. 5. ROI-based marginal profiles for frequency, spectral modulation, and temporal modulation. Each profile was obtained by averaging the MTFs along the other two dimensions and normalized (Norm.) by its maximum value. Based on the visual inspection of the full MTF (Fig. 4*B*) for the HG, PP, and PT, we computed distinct marginal tuning functions for frequencies below and above <2 kHz.

effect can occur, for example, if the activity of spatially separated neuronal populations is both specific to (combinations of) frequencies and modulations and time-dependent (10). However, many other neuronal and hemodynamic mechanisms likely contribute to our observations (34).

The ROI-based analyses revealed a number of interesting effects. In regions on the superior temporal plane (HG, PP, and PT), we observed a decrease in reconstruction accuracies for frequencies around 0.8 kHz, which corresponds to the frequency of peak energy for the scanner noise generated by our fMRI sequence (Fig. S5). In our clustered fMRI acquisition, sounds were presented during silent gaps between scans. It is thus possible that, similar to streaming paradigms (35), subjects modulated their attention to filter out the frequencies in the range of the scanner acoustic noise. Furthermore, the scanner noise between stimuli presentations might have affected the response to the auditory stimulation through, for example, adaptation of the neuronal population of interest or saturation of the BOLD response (36).

In the HG [the likely site of the primary auditory cortex (29)] and adjacent regions in the PP and PT, reconstruction accuracy for temporal modulation rates presented a clear dependency on frequency. For higher frequencies (>~2 kHz), accuracy was highest for faster modulation rates (6.6–30 Hz), which was not observed in the more lateral regions on the STG. For lower frequencies, the reconstruction accuracy profile for temporal modulation rates was highest in the range of 2–4 Hz, with a peak around 3 Hz. This finding is consistent with previous fMRI studies that examined cortical responses to temporal modulation rates with broadband noise (37, 38), and especially with those studies using narrow-band sounds (39, 40). However, these previous studies did not report the interdependency between sound carrier frequency and temporal modulations as observed in our study. Such interdependency may relate to the psychoacoustic observation that for higher frequency carriers, detection of temporal modulation changes is poorest at slower rates (41).

In STG regions, the reconstruction accuracy profiles for temporal modulations were highest in the range of 2–4 Hz, which is in agreement with most previous fMRI studies (37–40). Under the assumption that the observed reconstruction accuracy reflects the tuning properties of neuronal populations, our discriminability analysis suggested that such an amplification of acoustic components may lead to sound representations optimized for the fine-grained discrimination of speech (and voices) rather than for natural sounds per se (28). This interpretation is consistent with the hypothesis that neuronal populations in higher level auditory cortex are preferentially "tuned" to acoustic components relevant for the analysis of speech (23, 42). For human listeners, speech is arguably the class of natural sounds with the highest behavioral relevance, and it is thus reasonable that the human brain has developed mechanisms to analyze speech optimally. Importantly, the discriminability effects were significant already in the HG (Fig. S7). Furthermore, neither in primary regions nor in STG regions were the profiles of reconstruction accuracy for temporal modulations affected by the removal of speech/voice sounds (Fig. S6*B*). Together, these findings put forward the hypothesis that, in the human brain, even the properties of neuronal populations in early auditory cortical areas and the general purpose mechanisms involved in the analysis of any sound have been shaped by the characteristic acoustic properties of speech. This hypothesis is consistent with psychoacoustic investigations showing that human listeners have highest sensitivity in detecting temporal modulations changes in the range of 2–4 Hz, even when tested with broadband noise and tones (43, 44). Finally, the tight link between these mechanisms and speech predicts that these properties are specific to the human brain, which could be tested by performing the same (fMRI) experiments and analyses in nonhuman species.

Our study extends ECoG investigations on the representation of speech (12, 13, 35) to fMRI and to sounds other than speech. Although lacking the exquisite temporal resolution of ECoG, fMRI is noninvasive, enables large brain coverage, and approaches a spatial resolution in the submillimeter range (45). Together with the results of single-voxel encoding (14) (*SI Discussion*), the present study supports the hypothesis that the human auditory cortex analyzes the spectrotemporal content of complex sounds through frequency-specific modulation filters. Our proposed framework can be used to address relevant questions, for example, on how such processing changes due to ongoing task demands or to specific skill acquisitions (e.g., musical training, reading acquisition), brain development and aging, or hearing loss. Furthermore, in combination with submillimeter fMRI, it can be used to analyze the transformation of sound representations across cortical layers (45).

Whereas we choose to model the responses to real-life sounds, the MTF of a given region could be more simply estimated using synthetic sounds (e.g., dynamic ripples), with each one designed to include a unique combination of modulations (15). Compared with using synthetic stimuli, however, real-life stimuli appear advantageous for two reasons. First, they engage the auditory cortex in meaningful processing and, especially in nonprimary areas, they evoke larger responses compared with synthetic stimuli. Second, each stimulus contains a different combination of features of interest, such that the entire set of stimuli efficiently covers a wide range for each feature. The MTF of a region is then estimated by assessing which feature is accurately reconstructed from fMRI response patterns. This procedure allows estimating a region's reconstruction accuracy profile at the resolution of the representation model (discussed further in *SI Discussion*). With separately presented synthetic stimuli, obtaining such a resolution would involve the presentation of 15,360 different conditions, which is not practically feasible. Evidence from animal (46) and human (13, 47) electrophysiology indicates that spectrotemporal receptive fields and MTFs estimated using synthetic sounds are poor predictors of responses to natural sounds. In further studies, it would be relevant to investigate whether and to what extent this observation also applies to neuronal population responses as measured with fMRI.

Although efficient, the combination of real-life stimuli and model-based fMRI to study acoustic processing in auditory cortical regions has inherent caveats. Beyond the acoustic analysis, fMRI response patterns in the superior temporal cortex relate to higher levels of sound processing, as required, for example, for the perceptual and cognitive processing of speech (24, 25), voice (26, 48), or music (49, 50). These higher processing levels are not explicitly accounted for in the current sound representation model and may involve complex (nonlinear) transformations of the elementary acoustic features, which are the actual targets of the decoding. This simplified modeling may give rise to the possibility that a feature is successfully decoded only by virtue of its complex interrelation with other factors affecting the fMRI signal. The results of the analysis excluding speech and voice sounds confirmed that the tested model describes the processing of sounds in early auditory regions (HG, PT, and PP) as well as in STG regions well. These latter regions, however, were affected more by the exclusion of speech and voice sounds from the training set. Most likely, this finding reflects the fact that these regions are the sites of relevant (nonlinear) transformations of the acoustic input into higher level neural representations. The results of our discriminability analysis (Figs. S7 and S8) put forward the "zooming in" into an informative but limited subset of frequency/modulations as a mechanism that is potentially useful at the input stages of this transformation. Developing and testing computational descriptions of the full transformation chain, however, remain a challenge for future modeling and functional neuroimaging studies.

Materials and Methods

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. Anatomical MRI and fMRI data were collected at 7 T and preprocessed using BrainVoyager QX (Brain Innovations). Auditory spectrograms and modulation content of the stimuli, as well as fMRI-based reconstruction of spectrograms and waveforms, were obtained using "NSL Tools" (www.isr.umd.edu/Labs/NSL/Software.htm) and customized MATLAB code (The MathWorks, Inc.). Methods for estimating fMRI response patterns, for training and testing of the multivariate decoders, and for the statistical assessment of the results were developed and implemented using MATLAB (*SI Materials and Methods*).

ACKNOWLEDGMENTS. This work was supported by Maastricht University, the Dutch Province of Limburg, and the Netherlands Organization for Scientific Research (Grants 453-12-002 to E.F., 451-15-012 to M.M., and 864-13-012 to F.D.M.); the NIH (Grants P41 EB015894, P30 NS076408, and S10 RR26783); and the W. M. Keck Foundation.

1. Theunissen FE, Elie JE (2014) Neural processing of natural sounds. *Nat Rev Neurosci* 15:355–366.
2. Singh NC, Theunissen FE (2003) Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am* 114:3394–3411.
3. Fukushima M, Doyle AM, Mullarkey MP, Mishkin M, Averbeck BB (2015) Distributed acoustic cues for caller identity in macaque vocalization. *R Soc Open Sci* 2:150432.
4. Chi T, Gao Y, Guyton MC, Ru P, Shamma S (1999) Spectro-temporal modulation transfer functions and speech intelligibility. *J Acoust Soc Am* 106:2719–2732.
5. Drullman R, Festen JM, Plomp R (1994) Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am* 95:1053–1064.
6. Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. *PLOS Comput Biol* 5:e1000302.
7. Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. *Science* 270:303–304.
8. Rodríguez FA, Read HL, Escabí MA (2010) Spectral and temporal modulation tradeoff in the inferior colliculus. *J Neurophysiol* 103:887–903.
9. Miller LM, Escabí MA, Read HL, Schreiner CE (2001) Functional convergence of response properties in the auditory thalamocortical system. *Neuron* 32:151–160.
10. deCharms RC, Blake DT, Merzenich MM (1998) Optimizing sound features for cortical neurons. *Science* 280:1439–1443.
11. Kowalski N, Depireux DA, Shamma SA (1996) Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. *J Neurophysiol* 76:3524–3534.
12. Pasley BN, et al. (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251.
13. Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF (2016) Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. *J Neurosci* 36:2014–2026.
14. Santoro R, et al. (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLOS Comput Biol* 10:e1003412.
15. Schönwiesner M, Zatorre RJ (2009) Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc Natl Acad Sci USA* 106:14611–14616.
16. Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D (1991) Reading a neural code. *Science* 252:1854–1857.
17. Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J Neurophysiol* 102:3329–3339.
18. Miyawaki Y, et al. (2008) Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60:915–929.
19. Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. *Neuron* 63:902–915.
20. Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. *Nature* 452:352–355.
21. Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* 118:887–906.
22. Moerel M, et al. (2013) Processing of natural sounds: Characterization of multipeak spectral tuning in human auditory cortex. *J Neurosci* 33:11888–11898.
23. Poeppel D (2003) The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun* 41:245–255.
24. Binder JR, et al. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex* 10:512–528.
25. Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 18:903–911.
26. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403:309–312.
27. Woolley SM, Fremouw TE, Hsu A, Theunissen FE (2005) Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neurosci* 8:1371–1379.
28. Machens CK, Gollisch T, Kolesnikova O, Herz AVM (2005) Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron* 47:447–456.
29. Formisano E, et al. (2003) Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* 40:859–869.
30. 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.
31. Baumann S, et al. (2015) The topography of frequency and time representation in primate auditory cortices. *eLife* 4:4.
32. Barton B, Venezia JH, Saberi K, Hickok G, Brewer AA (2012) Orthogonal acoustic dimensions define auditory field maps in human cortex. *Proc Natl Acad Sci USA* 109:20738–20743.
33. Nishimoto S, et al. (2011) Reconstructing visual experiences from brain activity evoked by natural movies. *Curr Biol* 21:1641–1646.
34. Kriegeskorte N, Cusack R, Bandettini P (2010) How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? *Neuroimage* 49:1965–1976.
35. Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485:233–236.
36. Gaab N, Gabrieli JD, Glover GH (2007) Assessing the influence of scanner background noise on auditory processing. I. An fMRI study comparing three experimental designs with varying degrees of scanner noise. *Hum Brain Mapp* 28:703–720.
37. Giraud AL, et al. (2000) Representation of the temporal envelope of sounds in the human brain. *J Neurophysiol* 84:1588–1598.
38. Harms MP, Melcher JR (2002) Sound repetition rate in the human auditory pathway: Representations in the waveshape and amplitude of fMRI activation. *J Neurophysiol* 88:1433–1450.
39. Overath T, Zhang Y, Sanes DH, Poeppel D (2012) Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: fMRI evidence. *J Neurophysiol* 107:2042–2056.
40. Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nat Neurosci* 8:389–395.
41. Moore BC, Sek A (1995) Effects of carrier frequency, modulation rate, and modulation waveform on the detection of modulation and the discrimination of modulation type (amplitude modulation versus frequency modulation). *J Acoust Soc Am* 97:2468–2478.
42. Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: Emerging computational principles and operations. *Nat Neurosci* 15:511–517.
43. Bacon SP, Viemeister NF (1985) Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. *Audiology* 24:117–134.
44. Edwards E, Chang EF (2013) Syllabic (~2-5 Hz) and fluctuation (~1-10 Hz) ranges in speech and auditory processing. *Hear Res* 305:113–134.
45. De Martino F, et al. (2015) Frequency preference and attention effects across cortical depths in the human primary auditory cortex. *Proc Natl Acad Sci USA* 112:16036–16041.
46. Theunissen FE, Sen K, Doupe AJ (2000) Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *J Neurosci* 20:2315–2331.
47. Bitterman Y, Mukamel R, Malach R, Fried I, Nelken I (2008) Ultra-fine frequency tuning revealed in single neurons of human auditory cortex. *Nature* 451:197–201.
48. Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322:970–973 10.1126/science.1164318.
49. Zatorre RJ, Chen JL, Penhune VB (2007) When the brain plays music: Auditory-motor interactions in music perception and production. *Nat Rev Neurosci* 8:547–558.
50. Norman-Haignere S, Kanwisher NG, McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* 88:1281–1296.
51. Van de Moortele P-F, et al. (2009) T1 weighted brain images at 7 Tesla unbiased for Proton Density, T2* contrast and RF coil receive B1 sensitivity with simultaneous vessel visualization. *Neuroimage* 46:432–446.
52. 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.
53. Bishop CM (2006) *Pattern Recognition and Machine Learning* (Springer, New York).
54. Golub G, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21:9.
55. Kim JJ, et al. (2000) An MRI-based parcellation method for the temporal lobe. *Neuroimage* 11:271–288.
56. Good PI (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*, Springer Series in Statistics (Springer, New York).
57. Forman SD, et al. (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn Reson Med* 33:636–647.
58. Butts DA, Goldman MS (2006) Tuning curves, neuronal variability, and sensory coding. *PLoS Biol* 4:e92.