Progress in Neurobiology 200 (2021) 101982

Contents lists available at ScienceDirect

Progress in Neurobiology

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

Original Research Article

Cortical voice processing is grounded in elementary sound analyses for vocalization relevant sound patterns

Matthias Staib [a,*], Sascha Frühholz [a,b,c,d,*]

[a] Department of Psychology, University of Zurich, Zurich, 8050, Switzerland
[b] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, 8057, Switzerland
[c] Center for the Interdisciplinary Study of Language Evolution (ISLE), University of Zurich, 8057, Switzerland
[d] Department of Psychology, University of Oslo, Oslo, Norway

ARTICLE INFO

Keywords:
Voice
Auditory cognition
Auditory cortex
Communication
MVPA
fMRI

ABSTRACT

A subregion of the auditory cortex (AC) was proposed to selectively process voices. This selectivity of the temporal voice area (TVA) and its role in processing non-voice sounds however have remained elusive. For a better functional description of the TVA, we investigated its neural responses both to voice and non-voice sounds, and critically also to textural sound patterns (TSPs) that share basic features with natural sounds but that are perceptually very distant from voices. Listening to these TSPs, first, elicited activity in large subregions of the TVA, which was mainly driven by perpetual ratings of TSPs along a voice similarity scale. This similar TVA activity in response to TSPs might partially explain activation patterns typically observed during voice processing. Second, we reconstructed the TVA activity that is usually observed in voice processing with a linear combination of activation patterns from TSPs. An analysis of the reconstruction model weights demonstrated that the TVA similarly processes both natural voice and non-voice sounds as well as TSPs along their acoustic and perceptual features. The predominant factor in reconstructing the TVA pattern by TSPs were the perceptual voice similarity ratings. Third, a multi-voxel pattern analysis confirms that the TSPs contain sufficient sound information to explain TVA activity for voice processing. Altogether, rather than being restricted to higher-order voice processing only, the human "voice area" uses mechanisms to evaluate the perceptual and acoustic quality of non-voice sounds, and responds to the latter with a "voice-like" processing pattern when detecting some rudimentary perceptual similarity with voices.

1. Introduction

Previous research in human (Belin et al., 2000) and non-human primates (Petkov et al., 2008; Sadagopan et al., 2015) as well as in dogs (Andics et al., 2014) identified a region in the auditory cortex (AC) called the "temporal voice area" (TVA) (Belin et al., 2000). The TVA usually covers large parts of the AC and superior temporal cortex (STC) and has been defined by its higher neural activity in response to conspecific voices compared to other sounds. The TVA was recently divided into three anatomically distinct clusters (Pernet et al., 2015) located along the posterior-to-anterior direction of STC and has been characterized as being largely selective for voice processing with potentially only minor relevance for the processing of non-vocal auditory objects (Belin et al., 2018, 2000) and acoustically matched sounds (Agus et al., 2017). Given its proposed selectivity for voice processing, these clusters within the TVA were labelled as "voice patches" (Pernet et al., 2015), analogous to the face patches found in the visual system (Yovel and Belin, 2013). Despite these findings, voice selectivity in the cortically extended TVA as well as its voice patches is questionable for several reasons.

First, the overall extent of the TVA across many AC and STC subregions (Pernet et al., 2015) makes it rather unlikely that the entire area is reserved solely for voice processing, and some of these subregions in the primary and secondary AC are also involved in the acoustic analysis of other sounds (Leaver and Rauschecker, 2016). Second, the voice patches in the posterior STC (Belin et al., 2002, 2000; Kriegstein and Giraud, 2004) highly overlap with regions associated with social tasks and processing of visual stimuli (Allison et al., 2000; Blakemore, 2008; Isik et al., 2017), as well as tracking abstract parameters across sensory modalities (Schultz et al., 2005). Third, there are considerable

* Corresponding authors at: Department of Psychology, University of Zürich, Binzmuhlestrasse 14/18, 8850, Zürich, Switzerland.
E-mail addresses: matthias.staib@uzh.ch (M. Staib), sascha.fruehholz@uzh.ch (S. Frühholz).

https://doi.org/10.1016/j.pneurobio.2020.101982
Received 1 April 2020; Received in revised form 5 December 2020; Accepted 11 December 2020
Available online 15 December 2020

differences between voices and non-vocal sounds, not only in their perceptual quality, but also in higher-order acoustics and textural properties that could drive TVA activity. Hence, noisy and object-unlike sound patterns could elicit similar auditory cortical activity patterns (McDermott et al., 2011; Overath et al., 2015) as vocal sounds, albeit with an overall attenuated activity. This points to the possibility that, instead of a selectivity for voices, the TVA encodes biologically meaningful acoustic patterns prominent in, but not exclusive to, voices.

Accordingly, we aimed at a more basic mechanistic description of the functional brain activity in the TVA usually found for the neural processing of voices from a broader point of view. In the present study we therefore took a fundamentally different approach compared to previous studies that aimed at a more detailed description of the functional properties of the TVA. A previous study by Formisano and colleagues (Formisano et al., 2008) used an approach to manipulate the attentional focus of listeners to various types of voice information, such as voice identity and the semantic content carried by voices. It is one of the few studies in the field that used a multi-voxel pattern analysis approach, and found that different types of voice information is processed in overlapping and non-overlapping AC areas that are also covered by the TVA. However, because this study used voice stimuli only, it could not investigate the more basic functional properties of the TVA that lie underneath the level of voice processing. Other previous studies focused on creating synthetic or blended stimuli that acoustically mimicked entire voice sounds (Charest et al., 2013; DiMattina and Wang, 2006; Gentner and Margoliash, 2003; Toarmino et al., 2017) or important acoustic features of voices (Agus et al., 2017; Ghazanfar et al., 2007). The general aim of these previous studies was therefore to achieve a high similarity between original voice sounds and the synthetic counterparts, leading to the observation that neural responses are significantly higher to natural than to synthetic voice sounds (Gentner and Margoliash, 2003), at least in the right AC (Agus et al., 2017). While these previous studies might have provided a more differential picture on the supposed voice selectivity of some TVA subregions, they might be limited in providing a mechanistic description of the functional profile of the TVA that takes into account both acoustic sound features and especially perceptual dynamics in listeners. A second common limitation of these studies is the analysis of the overall TVA activation level only, which ignores information that the TVA might encode about non-vocal sounds observable only in multi voxel patterns.

Thus, instead of relying on acoustically matched, object-like sounds that mimic certain features of natural voices (Agus et al., 2017; Gentner and Margoliash, 2003; Ghazanfar et al., 2007), we investigated the commonly described TVA activity from a much broader acoustic space beyond acoustically matched synthetic sounds. Instead of introducing a high similarity of natural voices and synthetic sounds (i.e. high similarity in terms of acoustic matching of natural and synthetic voice sounds), we focused on achieving a high similarity of the neural activation pattern of the TVA by introducing completely object-dissimilar textural sounds patterns (TSPs) that are far from being perceived as natural voices. We investigated the potential of these TSPs to elicit a voice-similar TVA activity as elicited by natural sounds (i.e. high similarity on the neural activation side). If an object-dissimilar and voice-unlike sound can elicit the same TVA pattern as natural voices, this can provide more information about the functional processing properties of the AC regions underlying voice processing, with voice being a specific kind of auditory objects. Importantly, this pattern analysis decouples the information that is encoded in the TVA about non-voice sounds from the overall activity that the TVA expresses in response to voices or non-voices. The term “auditory object” here refers to an acoustic experience associated with source and event information about the object (Griffiths and Warren, 2004). We specifically aimed to test the hypothesis that the TVA does not exclusively respond to human voices as auditory objects, but is involved in processing more basic human as well as non-human sounds.

2. Methods

2.1. Participants

Twenty-five volunteers (14 female, mean age 26.4 years, SD = 4.96) participated in the fMRI experiment. Inclusion criteria were normal or corrected-to-normal vision and no history of neurological or psychiatric disorders. All participants gave written informed consent and were financially reimbursed for participation. The study was approved by the cantonal ethics committee of the Canton Zurich (Switzerland).

An additional group of 23 independent volunteers (14 female, mean age 26.5 years, SD = 3.56) were separately invited prior to the MRI experiment to rate the sounds but did not participate in the MRI experiment.

2.2. Stimuli and task

The set of natural sounds consisted of recordings of 70 vocalizations (speech and non-speech) and 70 non-vocalizations (animal, natural and artificial sounds) (Belin et al., 2000) of 500 ms duration. As described in (Belin et al., 2000), vocal stimuli were recorded from 47 speakers (from babies to elderly people) and were either speech (non-words) or non-speech (laughs, sighs, and various onomatopoeia). Non-vocal stimuli consisted of sounds from nature (e.g. wind, streams), animals (cries, gallops), the human environment (cars, telephones, planes) or musical instruments (bells, harp, and instrumental orchestra).

The set of TSPs was generated from modulated noise with the Gaussian Sound Synthesis Toolbox, Version 1.1 (McDermott et al., 2011). The intensity of all sounds was scaled to 70 dB using Praat (www.praat.org). The cochleagrams of TSPs have a multivariate, Gaussian-distributed, log-energy, time-frequency decomposition, in which the decay constants of the frequency and temporal correlation can be controlled, resulting in varying degrees of structure over time or over frequency (**Supplementary** Fig. 1). The structure of TSPs can mimic perceptual properties of natural sounds, but cannot unambiguously be assigned to any sound category, such as voices. The goal was to generate TSPs that perceptually range from TSPs that are perceived like voices to lesser or greater extents (i.e. a range of low to high "voice similarity"), but that are clearly distinct from human voices.

The selection process of the final 400 TSPs for the MRI experiment combined an analysis of acoustic features with an independent evaluation by listeners who did not participate in the fMRI study. First, 10′000 TSPs of 500 ms duration were generated, equally spaced on the frequency and temporal correlation parameters ranging from 0.01–2 (in steps of 0.02 for each parameter), resulting in 100 × 100 TSPs. For each TSP, 88 acoustic features (Eyben et al., 2016) were extracted with the publicly available toolbox openSmile v2.3.0 (Eyben et al., 2013). A subset of TSPs was then selected on the basis of a k-nearest neighbor (knn) classification of acoustic similarity to recorded vocal and non-vocal sounds. To this end, a knn classifier was fitted (Matlab function *fitcknn*) to separate the original voices from non-voice sounds in the space of the acoustic features and then used to predict the class membership for each of the 10′000 TSPs. The 600 TSPs closest to voices, and the 600 TSPs closest to non-vocal sounds were then rated by 23 independent volunteers outside the MRI scanner on a visual analogue scale according to their perceived voice similarity (**Supplementary** Fig. 2). These ratings were only used for the pre-selection and not used for any further analysis of the fMRI data. Finally, the 200 sounds with the lowest average voice similarity rating (between 2.0 and 17.7 %) and the 200 sounds with the highest average rating (between 37.0 and 57.9 %) were selected for the fMRI experiment. While these TSPs vary in their perceived voice similarity, none of them was mistaken for a real human voice. This was confirmed by an acoustical analysis and two additional ratings (Fig. 1**a-c**, **Supplementary** Fig. 1) obtained after the fMRI experiment was completed. Fig. 1a visualizes the acoustic overlap of TSPs with non-voices, shown here for the acoustic shimmer (the

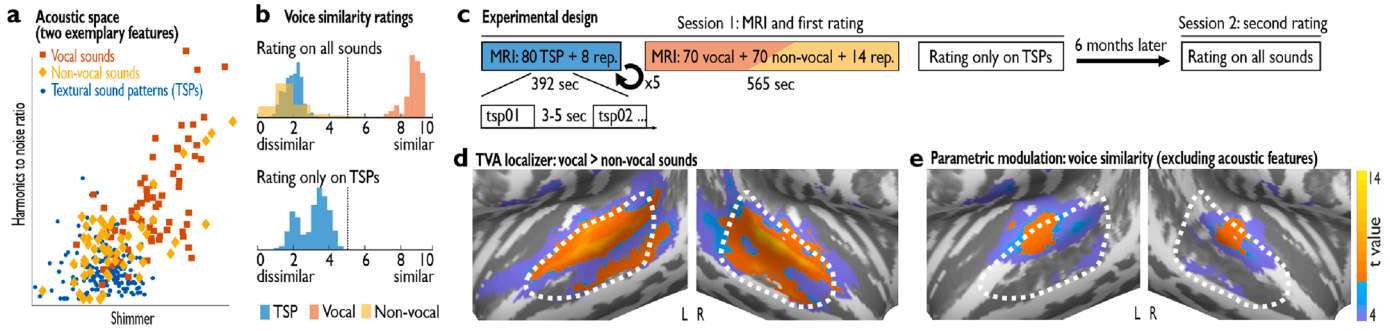

Fig. 1. The TVA processes perceptual features of voices and synthetic sounds.
(a) Textural sound patterns (TSPs, blue) differ acoustically from voices (red), exemplary shown for the amplitude variation of the sounds (shimmer), and their degree of acoustic periodicity (harmonics to noise ratio).
(b) Perceived voice similarity rated on all sounds (upper panel; n = 19) or on TSPs only (lower panel; n = 25).
(c) Participants were presented with 400 TSPs over 5 sessions in the fMRI experiment, followed by a TVA localizer session containing 70 voice and 70 non-vocal sounds in random order. Participants had to press a button to indicate identical consecutive sounds (one-back task). After the fMRI experiment was complete, participants rated the sounds outside the scanner, and again in a second session several weeks later.
(d) Group activation in bilateral AC for voice minus non-vocal sounds. White outlines depict anatomically defined TVA (see Methods). Blue, $p < 0.001$ (uncorrected), orange $p < 0.05$ (family-wise error corrected).
(e) Parametric modulation analysis of voice similarity of TSPs including 15 acoustic features as additional modulators.

amplitude variation of the sound), and the harmonics to noise ratio (degree of acoustic periodicity).

2.3. Acoustic differences of sound sets

The acoustic analysis of natural sounds (voice, non-voice sounds) and TSPs was based on 15 out of 88 (Eyben et al., 2016) acoustic features that we extracted with the toolbox openSmile (Eyben et al., 2013). Acoustic features were discarded if they could not be estimated for the majority of sounds (returning a value of zero), or if they highly inter-correlated with other features. The feature elimination was performed iteratively, removing in each step the feature that shows the highest occurrence of correlations above a threshold of $r = 0.8$ with the other features. The remaining acoustic features (Eyben et al., 2016) include:

Energy/Amplitude-related parameters:

- Loudness, estimate of perceived signal intensity from an auditory spectrum.

Spectral (balance) parameters:

- Spectral slope 0−500 Hz and 500−1500 Hz, linear regression slope of the logarithmic power spectrum within the two given bands.

Spectral (balance/shape/dynamics) parameters:

- MFCC 1–4 mel-frequency cepstral coefficients 1–4.
- Spectral flux difference of the spectra of two consecutive frames.

Arithmetic mean and coefficient of variation (standard deviation normalized by the arithmetic mean) were returned for all features (denoted as mean and SD in Fig. 2 and **Tab. S1**). For the loudness parameter, the 20th, 50th and 80th percentile, the range of 20th to 80th percentile, and the mean and standard deviation of the slope of rising/falling signal parts were also computed.

Supplementary Fig. 1 shows the acoustic properties of vocal and non-vocal sounds, together with all TSPs to supplement the analysis of acoustic confounds between voices and non-vocal sounds in the main text. Insets show that several acoustic features are correlated with rated voice similarity of TSPs.

To investigate the acoustic difference between voices and non-vocal sounds with a similar method as the classification scheme for brain patterns (main text), we used a support vector machine (SVM) to discriminate voices from non-vocal sounds by the 15 features in a 5-fold cross-validation scheme, producing a classification accuracy of 66.4 %. Next, we asked whether these 15 properties could discriminate TSP sounds from their rated voice similarity. We included only the 70 sounds with the highest and the 70 sounds with the lowest rated voice similarity, mirroring the classification analysis of fMRI data. The SVM was able to correctly predict these classes with 73.6 % accuracy (chance level at 50 %). We then tested whether the classification models share parameters by a cross-classification from natural sounds to TSPs. We found a classification accuracy of 32.9 %, i.e. far below chance, indicating a reversed relationship between voice labels in natural sounds and TSPs. This reversed relationship was also evident in the negative correlation of $r=-0.25$, $p < 0.003$, between the SVM posterior of a TSP sound being a vocalization and the rated voice similarity. Thus, the higher the rated voice similarity, the farther the acoustic similarity to real voices and vice versa. This result indicates that voice similarity of TSPs and natural sounds are acoustically not fully comparable. This finding is in line with the results presented in the main text, showing that a perceptual feature drives TVA activity over and above acoustic features, and that primary AC, which mainly analyses acoustic features, does not generalize between these sound sets.

2.4. Procedure

In the scanner, participants were listening to the sounds via OptoACTIVE headphones (Optoacoustics) with active noise cancellation that minimized scanner noise. For the fMRI experiment, the set of 400 TSPs was split into five blocks of 80 TSPs each (shuffled for each participant). Sounds were presented with an inter-stimulus interval between 3 and 5 s, with eight randomly chosen repetitions (10 %) of consecutive sounds in each block (Fig. 1c). A block started and ended with 20 s and 15 s of silence, respectively, resulting in a total block duration of 392 s. Participants in the MRI scanner were instructed to listen to the sounds and indicate with a button press when a sound matched the sound from one step earlier (one-back task). This task is orthogonal to the main analysis of the fMRI data (Agus et al., 2017) and was performed with high accuracy across conditions (**Supplementary** Fig. 3). The five blocks of TSP presentation were followed by a TVA localizer block in which 140 sounds (Belin et al., 2000) and 14 repetitions (randomly chosen from all TVA localizer sounds) were played with the same inter-stimulus interval and task (block duration, 656 s). The second presentation of each repeated sound was excluded from all fMRI analyses. Participants from

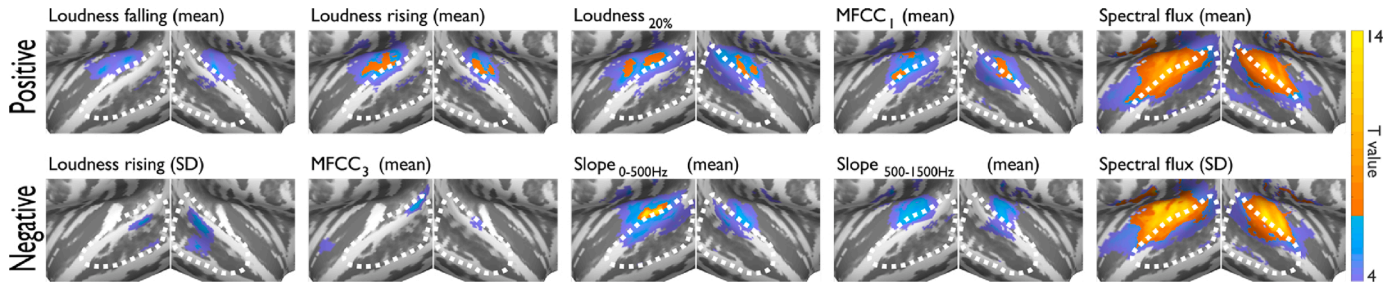

Fig. 2. Acoustic features, which vary across TSPs, modulate primary and higher AC activity.
Statistical analysis on the group level (n = 25, one-sample *t*-test) of activity modulation in bilateral AC by 10 of 15 tested acoustic features (Eyben et al., 2016) (see **Methods** and **Supplementary Information** for a detailed description). Displayed maps have at least one voxel above $t_{24=}|3.46|$ (p = 0.001, uncorrected). Upper panels: acoustic features that positively modulate voxel activation. Lower panels: acoustic features that negatively modulate voxel activation. Statistical maps show t-values, thresholded at p = 0.001 (blue), and family-wise error corrected at p = 0.05 (orange).
Abbreviations: MFCC mel-frequency cepstral coefficients.

the fMRI experiment rated the TSPs after completion of the MRI measurement and were invited again on a separate day several months later to evaluate the TSPs according to their perceived voice similarity and naturalness outside the MRI scanner.

2.5. fMRI data collection

Functional brain data were recorded in a 3T-Philips Ingenia with a standard 32-channel head coil. High-resolution structural MRI was acquired by using T1-weighted scans (field of view, 250 × 250 × 180.6 mm; matrix, 256 × 251; 301 1.20 mm overlapping sagittal slices). In each TSP block, 242 functional whole-brain images were recorded by using a T2*-weighted echo-planar pulse (EPI) sequence (TR 1.6 s, TE 30 ms, FA 82°; in-plane resolution 220 × 114.2 mm, voxel size 2.75 × 2.75 × 3.5 mm; gap 0.6 mm) covering the whole neocortex. For each participant, a whole-brain magnetic field mapping sequence (TR 30 ms, TEs 0.01/3.57 ms, FA 60°, voxel size 2 × 2.7 × 4 mm) was recorded to reduce image distortions from inhomogeneities in the magnetic field.

2.6. fMRI pre-processing

Pre-processing of fMRI data was performed by using standard procedures in the Statistical Parametric Mapping software (SPM12; Wellcome Trust Centre for Neuroimaging, London, UK; fil.ion.ucl.ac.uk/spm/software/spm12). Images were corrected for geometric distortions caused by susceptibility-induced field inhomogeneity (Cusack et al., 2003). A combined approach was used, which corrects for both static distortions and changes in these distortions from head motion (Andersson et al., 2001; Hutton et al., 2002). The static distortions were calculated for each subject from a B_0 field map that was processed by using the FieldMap toolbox as implemented in SPM12. With these parameters, functional images were then realigned and unwarped, a procedure that allows the measured static distortions to be included in the estimation of distortion changes associated with head motion. Slice time correction was performed to correct for differences in acquisition time of individual brain slices. No participant moved more than 2.9 mm in any direction during scanning. The motion-corrected images were then co-registered to the individual's anatomical T1 image by using a 12-parameter affine transformation. Deformation parameters for the transformation of atlas-based regions of interest (ROIs) in MNI space to the native space of participants' brains were obtained from a unified segmentation procedure as implemented in the Computational Anatomy Toolbox (CAT12; neuro.uni-jena.de/cat/).

2.7. Generation of the target TVA pattern

For the reconstruction analysis presented in the main text, we first computed, for each participant, the target activity pattern that represents the differences between presenting voices compared with non-vocal sounds. From the TVA localizer run, 70 vocalizations (speech and non-speech) were modelled as one regressor and 70 non-vocal sounds (artificial, animal and natural sounds) as a second regressor (Belin et al., 2000). We included, as regressors-of-no-interest, one regressor for the second presentation of each of the 14 sounds that were repeated (10 %) from a one-back task and a regressor for button presses. The fMRI model included six additional regressors of head motion that were estimated in the realignment step during pre-processing and 18 cardiac and respiration regressors (Kasper et al., 2017). The linear contrast of the first two regressors, [vocal minus non-vocal], served to identify voxels that are more sensitive to vocalizations than to non-vocal sounds.

2.8. Reconstruction of voice patterns

The aim of the pattern reconstruction was to reproduce the pattern obtained from the linear contrast [voice minus non-voice] of the TVA localizer by finding a weighted linear combination of activity patterns observed while participants listened to computer-generated TSPs. This approach tests our assumption that activity previously observed in the TVA based on a contrast of voices minus non-vocal sounds is not unique to the perception of voices, but that, instead, a set of synthetic sounds can result in similar activity patterns, modulated by the subjective evaluation of these synthetic sounds as vocalization. The information contained in the reconstruction model parameters was then related to the perceptual and acoustic quality of TSPs.

For the reconstruction, we first modelled each TSP trial in a separate general linear model (GLM) with one regressor for the current trial, one regressor for all other trials (least-squared, single trial (Mumford et al., 2012)) and the same set of regressors-of-no-interest as in the analysis for the TVA localizer. In a second step, the 400 resulting β-maps served as basis functions that were weighted to match the target pattern [voice minus non-voice] from the TVA localizer run. The reconstruction weights were computed with an L1 lasso regularization for a GLM regression. This method returns penalized maximum-likelihood fitted estimators for the linear model of the target pattern to the reconstruction patterns. This means that in addition to minimizing the deviance between the observed activity of the target pattern and the reconstructed activity, reconstruction weights of less informative patterns are shrunk to zero. Consequently, using the L1 norm for penalizing large weights favors sparse models. To control the penalty of large weights, we identified the optimal parameter λ of the penalty term in the optimization function through a 10-fold cross-validation. We constrained the model to return a maximum of 140 non-zero weights in order to limit model complexity, matching the generation of the target pattern.

Next, we tried to explain the information contained in the 140 reconstruction model parameters that are associated with 140 (out of

400) TSPs. Hypothetically, any set of randomly generated patterns could be used to fit the TVA brain pattern, but their model parameters would similarly be random and not entail any meaningful information. Instead, for the TSPs we tested whether their reconstruction weights follow the gradient of interpretable features (acoustically or perceptually), indicating a direct relation between the quality of the TSPs and the TVA voice processing pattern. Specifically, we hypothesized that for TSPs with a slightly higher rated voice similarity, a more positive reconstruction weight would be estimated, while for lower rated TSPs, a negative weight is more likely. This relation would follow the contrast weights for brain patterns of voices (+) and non-voices (-) in the computation of the TVA target pattern. We assumed that, on top of that, acoustic features might partly explain variance in the reconstruction weights. To directly test whether the perceptual rating can explain variance of the reconstruction weights over and above the acoustics, we defined a model with 16 fixed effects including rated voice similarity (full model),

$$\text{Weights} \sim \text{Acoustic feature}_1 + \ldots + \text{Acoustic feature}_{15} + \text{voice similarity} + \text{Error},$$

And, as baseline model, a nested model with 15 fixed effects (regressors-of-no-interest only),

$$\text{Weights} \sim \text{Acoustic feature}_1 + \ldots + \text{Acoustic feature}_{15} + \text{Error}.$$

Both models included random intercepts for participants and were estimated with the MATLAB function *glme*. A formal model comparison (MATLAB function *compare*) was conducted, based on the difference in model likelihood and complexity. The statistics of the likelihood ratio test can be approximated by a χ^2_1 distribution, with the degree of freedom as the difference in model complexity (16–15 = 1 parameter).

2.9. Parametric modulation analysis

For blood-oxygen level dependent (BOLD) responses of TSPs, we modelled each acoustic feature as a parametric modulator in a GLM with one regressor that included all sounds, one regressor for the acoustic property of interest and regressors-of-no-interest similar to the TVA localizer. We used a separate GLM for each parametric modulator to avoid effects of orthogonalization (Mumford et al., 2015). In addition, we created a model that included all 15 parametric modulators and one modulator for voice similarity (Fig. 1e).

2.10. Cross-prediction of voice similarity

For the decoding models, we trained a set of linear SVMs as implemented in The Decoding Toolbox, v3.96 (Hebart et al., 2015) (sites.google.com/site/tdtdecodingtoolbox/), to classify activity patterns from natural voices and non-voice sounds, from TSPs, and across natural sounds and TSPs (Fig. 4, **Supplementary Fig. 6**). This led to four classification schemes:

(i) natural → natural,
(ii) TSP → TSP
(iii) natural → TSP
(iv) TSP → natural

For these analyses, the TSPs were split into two groups, each containing 70 sounds with the highest rated voice similarity, and the 70 lowest, respectively. The remaining TSPs were discarded to match the number of 140 natural sounds. Classification schemes i-ii entailed a 5-fold cross-validated classification in which the model was trained on 112 out of the 140 trials and tested on the remaining 28 trials. This was repeated such that each trial served as a training data point four times and as a test data point once. Classification schemes iii-iv used cross-classification, in which all 140 trials from one sound set were used for training and the 140 trials from the other set used for testing. For the classification, activity patterns from natural voices/non-vocal sounds were modelled as single trials, similar to the GLM described for TSPs, and each trial was labelled as voice (+1) or as non-voice (-1). For analysis iii, i.e. cross-classification from natural to TSP, we additionally directly compared this SVM decision value with rated voice similarity. This analysis has the advantage that informative variance of rated voice similarity is not lost during assignment of binary labels. We again used a model comparison similar to the analysis described above. A full model that included 15 acoustic properties and voice similarity was tested against the nested model without voice similarity.

2.11. Statistical testing of decoding accuracies based on prevalence inference

To evaluate the significance of classification accuracies, we performed a permutation-based non-parametric test. It has been argued that in the case of information-like outcomes, such as classification accuracies, classic statistical tests (including one-sample t-tests) do not conform to random effect testing, which is required for valid generalization to the population (Allefeld et al., 2016). This is based on the fact that the true (unobservable) classification accuracy of class membership (e.g. vocal and non-vocal sound) can never be below the chance level of 50 %, because the quantity of information can never be negative. This stands in contrast to classical tests of brain activation amplitudes of the BOLD signal, which have a normally distributed error, allowing for a negative error. It follows that for information-like measures, such as classification accuracies, the true (unobservable) population null distribution cannot be described as a normal distribution with a mean at the chance level. The constraint specifying that any true accuracy is equal to or greater than chance level (but never smaller) implies that an observed accuracy distribution around zero can only arise if all accuracies are truly zero plus a normal distributed error (which can be negative). A negative error (resulting in an observed below-chance accuracy) can arise when assumptions are not met, such as the test and training data sets in a cross-validation scheme coming from non-identical distributions. Again, this contrasts with other outcomes in which the true values can be negative and therefore the null distribution arises from positive and negative true values, as is known, for example, from BOLD estimates. For decoding accuracies, this implies in turn that above-chance group results will arise immediately if at least one participant has above-chance accuracy, as this value cannot be outweighed by the observation of a (truly) negative accuracy from another participant. Conceptually, this situation describes a fixed-effect analysis in which statistical results are suitable to describe the observed sample but do not generalize to the population.

To implement these statistical considerations in our analysis, we used the publicly available MATLAB toolbox Prevalence-Permutation (github.com/allefeld/prevalence-permutation/) to compute permutation-based information prevalence inference with the minimum statistic. This method derives group statistics for “information-like” data in which underlying values are assumed to be strictly positive. Specifically, for each decoding scheme, participant and ROI, we computed the decoding accuracy by using the correct class labels "vocal" and "non-vocal" and 1000 of $P_1 = 2^{70-1}$ unique permutations. For information prevalence inference, the permutations of $N = 25$ participants were combined. We randomly selected $P_2 = 10^7$ of $P_1^N = 1,000^{25}$ possible combined permutations at the group level (Allefeld et al., 2016). The distribution of accuracies after randomly permuting the class labels is subject to the same constraints (strictly positive true accuracies plus normally distributed error) as the accuracy obtained from the correct class labels. The prevalence inference method allows testing of two hypotheses: the population null hypothesis and the majority null hypothesis. The population hypothesis asks whether classification is above chance for anyone in the population, whereas the majority hypothesis

tests whether classification accuracy is above chance for the majority of the population.

2.12. Definition of ROIs

We created a functionally defined region-of-interest (ROI) for each participant's brain, including only voxels in the temporal cortex that responded to any natural sounds (including voices and non-vocal sounds) with a liberal threshold of $p = 0.01$ (uncorrected). To investigate regional differences within the auditory cortex, we defined a set of atlas-based anatomical ROIs, including primary (Te1.0, Te1.1, Te1.2) (Morosan et al., 2001), secondary (BA42) (Van Essen and Dierker, 2007) and higher auditory regions (Te3, MTG, STG) (Morosan et al., 2005; Tzourio-Mazoyer et al., 2002), that were warped into the native space of each participant (**Tab. S2**).

Our main analyses focus on the anatomically defined region Te3 to represent the TVA, because a careful inspection on TVA literature shows inconsistent definitions and sizes across studies (Aglieri et al., 2018; Agus et al., 2017; Belin et al., 2000; Pernet et al., 2015). A functional definition of the TVA (a) strongly depends on the power of the design, (b) relies on decisions from the researcher about the constraints (significance threshold), (c) biases multivariate methods by only including significant voxels of mass-univariate analyses, (d) potentially introduces double-dipping to the classification analysis of vocal vs. non-vocal, (e) partly overlaps with primary areas, and (f) introduces hemispheric asymmetries which might lead to spurious effects. Overall, this classical TVA definition is essentially derived from mass-univariate methods which we aimed to overcome. As an alternative we base our analyses on the Te3, because it contains major proportions of the TVA, is an unbiased and reproducible ROI, and is similar across hemispheres.

All mass-univariate results are shown for the full field-of-view (entire brain), without any masking.

3. Results

3.1. Acoustic and perceptual properties of non-vocal synthetic sounds

A sound set of 400 synthetic non-vocal sounds (i.e. TSPs) were selected for the MRI experiment (see **Methods**), such that they not only varied in several acoustic features similarly to human voices and non-voices (**Supplementary** Fig. 1b) but covered a broad range of perceptual qualities. To ensure that the TSPs are perceptually clearly distinct from human voices, we asked the participants after completion of the MRI experiment (Fig. 1c) to rate the perceptual similarity of each TSP to the human voice in a range between 1–10, where "1″ denoted strongest voice-dissimilarity, and "10″ denoted highest voice-similarity (Fig. 1b, bottom). In a second follow-up session, the same participants ($n = 19$, due to 6 dropouts) were invited again to rate the TSPs together with voices and non-vocal sounds (Fig. 1b, top). Ratings for TSPs were highly overlapping with the ratings for non-vocal sounds and completely distinct from voices. The highest rating for any of the 400 TSPs was 3.42 ($SD = 2.51$). This rating is comparable to the non-vocal sound "animal_12.wav" (Capilla et al. 2013) which has a rating of 3.52 ($SD = 2.13$). Thus, our set of TSPs met the critical requirement to independently test TVA responses to acoustic features in the absence of voices and voice perception.

To understand the acoustic foundation of the voice similarity ratings, we related them to 15 spectral and energy-related features (**Supplementary** Fig. 1, see also **Methods** for details on acoustic feature selection). A regularized regression model explained 38.9 % variance of rated voice similarity (based on 12 out of 15 features). We similarly tried to explain the ratings of the original voices and non-voices (Fig. 1b) with their acoustic features and found that 35 % of their variance could be explained on 12 features. Of these features, only 6 were shared between the models (**Supplementary Tab. 1**). The rated voice similarity therefore signifies an abstract feature that could explain voice perception over and above the acoustic features tested here.

3.2. Representation of non-voice perception in the TVA

The classical TVA localizer experiment (Belin et al., 2000) identified voice-sensitive brain regions by contrasting activations elicited by voices against non-vocal sounds, which we replicated here (Fig. 1d). This showed bilateral activity in the primary and non-primary auditory cortex (AC), classically referred to as a specialized voice-sensitive cortical area, the TVA (Belin et al., 2000).

To investigate whether responses of individual voxels in this area track perceptual and acoustic properties of both voices and synthetic sounds, a parametric modulation analysis of TSPs (Fig. 1e) was performed to test whether the rated voice similarity drives activation in the TVA. For this model, all 15 acoustics features were included as additional parametric modulators. The continuous response of the AC to these acoustic features was then assessed individually (Fig. 2), showing that several loudness- and spectrum-related acoustic features of TSPs, such as spectral flux, previously associated with voice perception (Overath et al., 2008), activate primary AC regions, but partly extend into the original TVA.

Thus, variance in the perceived voice similarity and acoustic features of TSPs might partly drive TVA activity in single voxels, indicating that this region might be sensitive to a variety of sounds that share some acoustic patterns with natural sounds.

3.3. Brain activations of voice-sensitivity can be reconstructed with synthetic sounds

The main goal of the study was to show that the typical multi-voxel brain pattern of voice processing (as compared to processing of non-voice sounds) in the TVA carries information about non-voice sounds. To this end, we introduce a reconstruction approach where the neural activity pattern of voice processing in the TVA is linearly modelled with activation patterns from TSPs (Fig. 3). For these multi-voxel pattern analyses, we defined the TVA as the anatomical region Te3 (Morosan et al., 2005) of the AC which showed the largest overlap with the functional activation of this and previous studies (Agus et al., 2017; Ahrens et al., 2014; Belin et al., 2000; Pernet et al., 2015). This definition allows an independent analysis of TVA activation patterns from voice-processing and is reproducible across participants (see **Methods** for a detailed ROI definition and discussion).

Using a regularized (L1) regression model, a weighted linear combination of brain activity patterns from the TSPs (Fig. 3b) was computed to estimate the TVA activation pattern, separately for each hemisphere and participant. Regularization reduces model complexity and thereby avoids overfitting to noise in the signal. As an additional model constraint, a maximum of 140 non-zero weights were allowed. On average, 110 brain patterns of TSPs returned with a non-zero weight which was sufficient to reconstruct the TVA pattern with $R^2 > 0.9$ for either hemisphere (Fig. 3b). A comparison of Fig. 3**a-b** shows that a similar pattern of TVA activity can indeed be reconstructed by TSPs that are not perceived as human voices, non-vocal sounds, or any other auditory object. The reconstruction modelling was also extended to adjacent ROIs outside TVA, as detailed in the Supplementary Information (**Supplementary** Fig. 4), showing that the effects are characteristic of higher-level but not primary AC.

To verify that our reconstruction model captured signal that is related to voice perception, we evaluated the reconstruction weights by three methods. First, TSPs consistently associated with positively, as compared to negatively weighted patterns across participants elicited AC activity that resembles the original TVA activity (**Supplementary** Fig. 4a). Second, we tested consistency of the weights by reconstructing the neural patterns from the contralateral hemisphere, resulting in a correlation between the reconstructed and true TVA pattern of $r = 0.33$ from left to right, and $r = 0.37$ for the other direction (Fisher-z

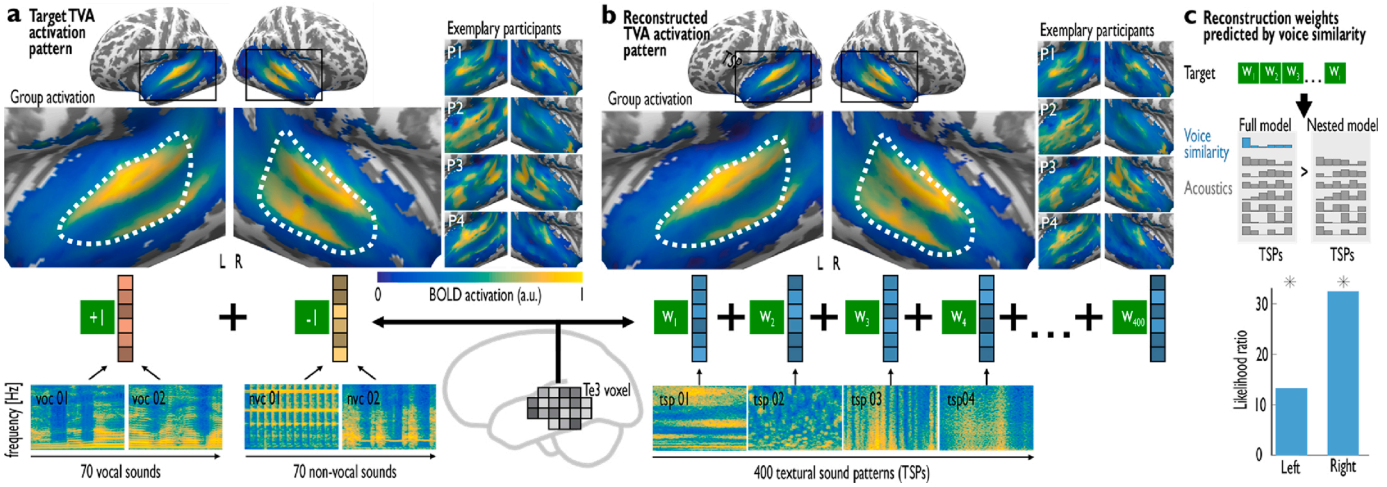

Fig. 3. Activation elicited by natural voices can be reconstructed from synthetic sounds.
(a) Normalized group level AC activations of voices (+1) minus non-vocal sounds (-1) (dashed line shows TVA).
(b) For each participant, the target pattern in (a) was reconstructed from multivoxel brain patterns of 400 TSPs. Per hemisphere, a reconstruction weight (w_1, … w_{400}) was estimated for each TSP using a regularized regression model. The surface plot includes reconstructions of regions outside the TVA (see **Supplementary Information**).
(c) Comparison of two linear models predicting the reconstruction weights either by acoustic features and rated voice similarity (full model) or acoustic features only (nested model). Bars show test statistics of a likelihood ratio test (χ^2_1 distributed) for the for the full model over the nested model. * $p < 0.05$; Bonferroni-corrected for two hemispheres.
Abbreviations: P(1–4) participants 1–4; L left; R right; Te temporal regions; TVA temporal voice area.

transformed and averaged across subjects). This confirms that the reconstruction weights contain some information about the TSPs, partly consistent across hemispheres. To put this into perspective, the reconstruction weights from a control area (bilateral occipital cortex) were computed and entered into the reconstruction model of the TVA, resulting in average correlations of $r = 0.14$ and $r = 0.16$, respectively, i.e. significantly less than for the contralateral TVA (paired *t*-test of Fisher-z transformed correlation coefficients for the TVA and control area, $t_{24} = 3.30$, $p = 0.003$ from left to right, and $t_{24} = 4.83$, $p = 0.0001$ from right to left, respectively).

Finally, we directly associated the reconstruction weights with the perceptual and acoustic features (Fig. 3c) by testing whether the reconstruction weights could be explained equally well by the acoustic features of TSPs, or whether the perceptual feature of voice similarity can explain a unique proportion of the weights variance (Fig. 3c). This directly tests whether the reconstruction of the TVA pattern follows the gradient of features from non-voice TSPs. To do so, we modelled the reconstruction weights across all participants with the acoustic features together with the rated voice similarity (full model) using a mixed-effects model with the participant as random intercept (see **Methods** for the formal model definition). A nested model, excluding rated voice similarity (but including all acoustic features) was estimated as baseline model to quantify the contribution of the ratings to the prediction of reconstruction weights of brain activity patterns. A formal model comparison was conducted (see **Methods**), based on the differences in model likelihood (χ^2_1 distributed) and model complexity, indicating that the ratings indeed explain variance of the reconstruction weights over and above acoustic features for the left ($\chi^2_1 = 13.3$, $p = 2.6 \times 10^{-4}$) and right TVA ($\chi^2_1 = 32.5$, $p = 1.2 \times 10^{-8}$). As shown in **Supplementary** Fig. 4 and in the **Methods**, this was also true for AC regions extending across the middle temporal gyrus, superior temporal gyrus, and areas covered by the original TVA, but not for primary AC. Consistently, the reconstruction weights correlated significantly with the ratings (**Supplementary Fig. 5**). Overall, these results show that the reconstruction of voice activity in TSPs partly follows subjective perceptual ratings, indicating that the TVA extracts voice-affine information hidden in structured noise stimuli, and does not exclusively respond to generic voice signals.

3.4. A common representation of voices and non-vocal sounds

To confirm these results in a complementary approach, we aimed to determine the common neural information that drives AC activity in response to both human voices and to TSPs by using multi-voxel pattern classification (Fig. 4, **Supplementary Fig. 6**). In a 5-fold cross-validation scheme (Fig. 4a), we trained a classifier to discriminate brain patterns of voices from patterns of non-vocal sounds. In the cross-validation scheme, 80 % of the data served as training data, and the remaining activity patterns as test data, repeated 5 times, such that each activity pattern served as test data once. The statistical significance of the classification accuracies was computed with a permutation-based non-parametric test (Allefeld et al., 2016) (see **Methods**) and p-values were corrected for the two tested brain regions (left and right TVA). Of the unseen brain patterns the classifier correctly identified 66.3 % (in left TVA) and 66.4 % (right TVA) as originating from the presentation of voices or non-vocal sounds, with $p = 2 \times 10^{-7}$ in both hemispheres (chance level 50 %).

To apply a comparable binary classification scheme to the TSPs based on their rated voice similarity (Fig. 4b), each TSP brain pattern needed to be assigned to a class, which does not directly follow from the continuous rating. To accommodate for that, we included only the 70 TSPs with the highest ratings against the 70 TSPs with the lowest ratings, and trained and tested on brain activity patterns from these two classes with the same cross-validation scheme as for voices and non-vocal sounds. Test accuracies of highest and lowest rated TSPs were 61.0 % ($p = 0.005$) in left TVA, and 58.9 % ($p = 0.0002$) in right TVA, respectively, confirming that the TVA distinguishes not only voices from non-vocal sounds, but also discriminates more or less voice-similar synthetic sounds that are overall clearly different from human voices. This is consistent with a classification of the TSPs acoustics (see **Methods**), showing that highest and lowest rated TSPs differ acoustically as well.

To answer whether distinguishing highest from lowest TSPs in the TVA generalizes to voices and non-vocal sounds, i.e. determining the common neural representation between these sound sets, we used a cross-classification approach between voices/non-voices and TSPs (Fig. 4c). For this, brain patterns from the lowest/highest TSP were used

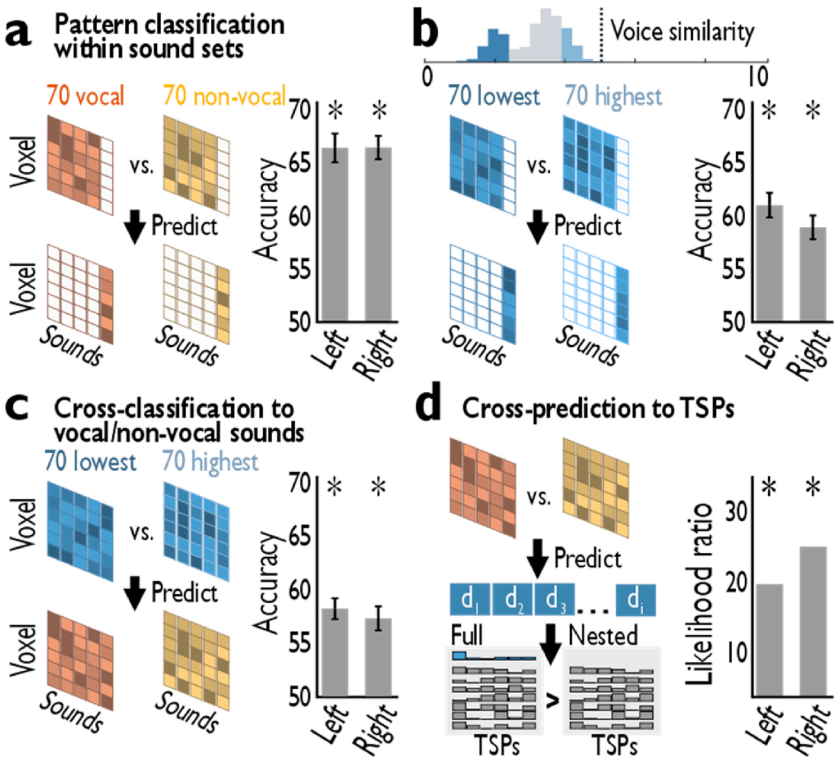

Fig. 4. A common neural representation of voice similarity across acoustic domains.
(a) Pattern classification of voice sounds (red) against non-voice sounds (yellow) in a 5-fold cross-validation scheme. Accuracy is shown for unseen test data. * $p < 0.05$ (prevalence-inference (Allefeld et al., 2016)); Bonferroni-corrected for two hemispheres. Error bars show +/- s.e.m across participants.
(b) Brain patterns of the 70 sounds with the highest voice similarity rating were classified against the 70 lowest rated sounds.
(c) Cross-classification from TSPs to voices and non-vocal sounds. The decoding scheme is similar to that in (b), but all 140 TSPs served as a training data set and all 70 voices and 70 non-vocal sounds as a test set.
(d) Cross-prediction of the support vector machine decision values (d_1, … d_i), i.e. predicted voice similarity, by either voice similarity and 15 acoustic features (full model), or acoustic features only (nested model). Bars show test statistics of a likelihood ratio test (χ^2_1 distributed). * $p < 0.05$; Bonferroni-corrected for two hemispheres.

to train a classifier, and the patterns from voices and non-vocal sounds served as test set. We found above-chance cross-classification accuracy in bilateral TVA (58.2 %, $p = 0.0009$ for left, and 57.2 %, $p = 0.0009$ for right), showing that the mechanism of distinguishing highest from lowest TSPs is partly shared with voices and non-vocal sounds.

The accuracy of the opposing cross-classification direction from natural voices and non-vocal sounds to TSPs in turn did not exceed 50 % (**Supplementary Fig. 6**), because the classifier trained on voices and non-vocal sounds was biased towards identifying TSP patterns as non-vocal sounds. This directly reflects the perceptual rating of TSPs as highly voice-dissimilar. To analyze whether any information from voices and non-vocal sounds generalizes to TSPs, we directly explained the classifier's continuous decision value with the rated voice similarity of TSPs (Fig. 4d), omitting to introduce an artificial dichotomy between high and low TSPs. This decision value represents the likelihood of a brain pattern to be more similar to either a voice or non-voice brain pattern and thereby closely resembles a predicted voice similarity for each TSP. Thus, this cross-prediction approach shows mutual information between predicted and rated voice similarity of TSPs, independent of a bias. Two models that explain the classifier's predicted voice similarity were then compared, i.e. using either rated voice similarity combined with acoustic features (full model), or by the acoustic features only (nested model), analogous to the reconstruction analysis introduced in Fig. 3**c**. A model comparison showed that the full model can explain the predicted voice similarity of TSPs over and above the nested model in the left ($\chi^2_1 = 19.1$, $p = 1.3 \times 10^{-5}$) and right TVA ($\chi^2_1 = 24.4$, $p = 7.9 \times 10^{-7}$), indicating that the predicted voice similarity of TSPs, obtained from a training on natural voices and non-vocal sounds directly relates to the voice similarity rated by participants. These cross-prediction analyses demonstrate that the TVA codes the differences between voices and non-vocal sounds with similar activation patterns as the gradient of perceived voice similarity of TSPs.

3.5. Analysis in primary and higher auditory cortex

We extended the reconstruction analysis and pattern classification to brain regions outside the TVA to investigate difference between primary and higher AC (**Supplementary** Fig. 4, **Supplementary Fig. 6**). That way we were able to show that most of the reported effects are characteristic to higher, but not primary AC. For the reconstruction approach, the TVA, MTG, STG and the functionally defined sound-sensitive region showed a significant relation between the reconstruction weights and rated voice similarity over and above acoustic features. The classifier significantly predicted multivoxel brain patterns as originating from the presentation of natural vocal or non-vocal sounds in several higher AC ROIs, with the highest accuracy of 68.8 % ($p = 10^{-7}$) in right sound sensitive voxels, averaged across participants. Across participants, classification of voice-similar TSPs versus voice-dissimilar TSPs was significantly above chance in several higher AC regions (**Supplementary Fig. 6**), and in the left Te1.0 (55.3 %, $p = 0.003$). For primary areas, we found that in right Te1.0, classification was significantly above chance for TSP with lowest vs. highest voice similarity (55.3 %, $p = 0.0029$). No other classification schemes were significant in Te1.0, Te1.1, or Te1.2, showing that pattern discrimination of voices and non-vocal sounds, highest vs. lowest TSPs, and generalization between these sound sets is mostly constrained to higher AC regions.

4. Discussion

Recent studies in humans (Agus et al., 2017; Belin et al., 2018) and nonhuman primates (Gil-Da-Costa et al., 2006; Petkov et al., 2008) proposed a specialized brain region in the AC was supposed to selectively responds to voices compared to other sounds. However, there is some reason to believe that this region is much less voice-selective than previously assumed. Here, we first showed that that processing of some complex acoustic features might extend into TVA, providing indication

that some subregions of the TVA respond to the acoustic patterns (i.e. energy- and amplitude-related parameters, spectral slope, spectral flux, and MFCC) of non-voice sounds (Herdener et al., 2013; Hullett et al., 2016). For the sounds presented in the present study, spectral flux elicited the largest extended activations in the AC and large parts of the TVA, with positive effects of the mean spectral flux, and negative effects of the spectral flux variation across the sound. Thus, in the TVA, there might exist a more general representation of biologically relevant acoustic information encoded in sound features that are common both to TSPs and natural voices. The validity of data-driven approaches that focus on the acoustic features of speech-related sounds to explain neuronal responses in the AC was recently demonstrated in a magnetoencephalography study, showing that simple acoustic models can be superior to encoding models that are based on higher-order constructs (Daube et al., 2019).

Next, we went one step further by demonstrating that the spatial profile of brain patterns of voice processing in the TVA can be reconstructed with high accuracy from activations by these TSPs. We focused on the Te3 (excluding primary AC), which is the predominant auditory cortical region of the voice patch system (Belin et al., 2018; Pernet et al., 2015), but seems to evaluate auditory patterns on a more fundamental level, related to the spectro-temporal texture across many sounds (Theunissen and Elie, 2014). The subjective perception of the TSPs significantly determined the reconstruction model, most evident in bilateral higher-level AC, with some effects in right primary AC. This suggests that, especially for the higher-level AC, the link between the TVA pattern and the reconstruction by the TSPs is partly based on a perceptual feature registered by listeners that cannot be entirely reduced to simple acoustics. Importantly, this link is not observable through the overall activation amplitude of the TVA, but the generalization of the activation pattern observed during voice processing. The underlying function of the TVA activity could therefore be the tracking of the perceptual quality of voice similarity even for non-vocal sounds and simple sound patterns (Hausfeld et al., 2018; Webster et al., 2017). This is further evidence that the TVA processing mechanisms generalize to basic sound evaluations and perceptual discrimination of non-vocal acoustic patterns, and are not selectively reserved for high-level object recognition (Santoro et al., 2014).

A limitation to our analyses involving basic acoustics directly follows from the selection strategy of the TSPs which is based on their perceptual quality, but not their acoustic features. This selection precludes a perfect acoustic match between the synthetic and the natural sound sets and might therefore underestimate the role of some acoustic features in the TVA.

Finally, we used a complementary multivariate classification analysis to demonstrate a common representation of acoustic processing in the TVA across sound sets. We first confirmed that natural voices and non-vocal sounds are neurally distinct in the TVA and other higher-level AC regions. This pattern of results is largely similar to the neural distinction between TSPs with the highest against the lowest voice similarity rating. Critically, in a cross-classification approach we found that a classifier trained on distinguishing high from low-voice-similar TSPs represents all necessary information to distinguish voices from non-vocal sounds in the TVA. In the reverse direction, we found that a classifier was able to predict neural activity for TSPs on a continuous basis in the TVA. In low-level auditory regions, no cross-classification effects were found, which is in line with our reconstruction approach, suggesting that the primary AC predominantly represents more basic acoustic properties that largely differ between natural sounds and TSPs, and can therefore not consistently generalize across these two acoustic domains.

Our method to test processing models of higher AC by designing stimuli based on a perceptual model, instead of presenting acoustic equivalents of voices, is a promising approach. For example, (Norman-Haignere and McDermott, 2018) have shown that primary AC is primarily driven by the acoustic profile of sounds whereas higher areas encode increasingly abstract representations of sounds. In our study, the highest level of abstraction that we tested is perceived voice-similarity that drives the TVA but not primary AC. Yet, in our study parts of the TVA simultaneously represents some basic acoustic properties of the TSPs. It is therefore likely that the representation of acoustic features is maintained even in higher AC unless it is superseded by higher-order properties of incoming sounds that are associated with abstract concepts.

It is important to note that our results only capture the categorical distinction between processing human voices and non-voice sounds, and do not challenge findings of the TVAs processing mechanisms for voices and for specific social information carried by voice signals itself. While our results show that the TVAs sound processing extends beyond voices, they are fully consistent with its ability to encode other vocal features and voice information, such as voice identity (Perrodin et al., 2015).

In summary, unlike the assumed functional specialization (Pernet et al., 2015; Petkov et al., 2008) and even functional selectivity (Belin et al., 2000) of many subregions of the AC for voice processing, the human TVA, or parts of it, seems to be more generally involved in processing a broad variety of acoustic patterns and potentially represents some measure for a sound to originate from a living source or its relevance for communication. It could be speculated that this generalization of the voice processing activation pattern to non-voices serves as a preparatory mechanism that pre-evaluates and filters a variety of sounds before the TVA engages in voice processing that is typically associated with an overall increase in activity. Alternatively, it is possible that the functional specialization of the AC for processing voices is less pronounced than frequently proposed and instead suggests a larger degree of functional flexibility with regards to sound object processing and sound feature sensitivity (Bandyopadhyay et al., 2010; Rothschild et al., 2010; Yildiz et al., 2016). Such flexibility and broader sensitivity to natural sound patterns seems evolutionarily more plausible than extended functional specialization across large cortical areas.

Author contributions

M.S. and S.F. contributed to designing the experiment, data acquisition, data analysis and writing the manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgments

The study was supported by the Swiss National Science Foundation (SNSF PP00P1_157409/1 and PP00P1_183711/1 to SF).

Appendix A. The Peer Review Overview and Supplementary data

The Peer Review Overview and Supplementary data associated with this article can be found in the online version: https://doi.org/10.1016/j.pneurobio.2020.101982.

References

Aglieri, V., Chaminade, T., Takerkart, S., P, B, 2018. Functional connectivity within the voice perception network and its behavioural relevance. Neuroimage 183, 356–365. https://doi.org/10.1016/j.neuroimage.2018.08.011.

Agus, T.R., Paquette, S., Suied, C., Pressnitzer, D., Belin, P., 2017. Voice selectivity in the temporal voice area despite matched low-level acoustic cues. Sci. Rep. 7, 11526. https://doi.org/10.1038/s41598-017-11684-1.

Ahrens, M.M., Shiekh Hasan, B.A., Giordano, B.L., Belin, P., 2014. Gender differences in the temporal voice areas. Front. Neurosci. 8, 228. https://doi.org/10.3389/fnins.2014.00228.

Allefeld, C., Görgen, K., Haynes, J.D., 2016. Valid population inference for information-based imaging: From the second-level *t*-test to prevalence inference. Neuroimage 141, 378–392. https://doi.org/10.1016/j.neuroimage.2016.07.040.
Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cogn. Sci. 4, 267–278. https://doi.org/10.1016/S1364-6613(00)01501-1.
Andersson, J.L.R., Hutton, C., Ashburner, J., Turner, R., Friston, K., 2001. Modeling geometric deformations in EPI time series. Neuroimage 13, 903–919. https://doi.org/10.1006/nimg.2001.0746.
Andics, A., Gácsi, M., Faragó, T., Kis, A., Miklósi, Á., 2014. Voice-sensitive regions in the dog and human brain are revealed by comparative fMRI. Curr. Biol. 24, 574–578. https://doi.org/10.1016/j.cub.2014.01.058.
Bandyopadhyay, S., Shamma, S.A., Kanold, P.O., 2010. Dichotomy of functional organization in the mouse auditory cortex. Nat. Neurosci. 13, 361–368. https://doi.org/10.1038/nn.2490.
Belin, P., Zatorre, R.J., Lafallie, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312. https://doi.org/10.1038/35002078.
Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Cogn. Brain Res. 13, 17–26. https://doi.org/10.1016/S0926-6410(01)00084-2.
Belin, P., Bodin, C., Aglieri, V., 2018. A "voice patch" system in the primate brain for processing vocal information? Hear. Res. 366, 65–74. https://doi.org/10.1016/j.heares.2018.04.010.
Blakemore, S.J., 2008. The social brain in adolescence. Nat. Rev. Neurosci. 9, 267–277. https://doi.org/10.1038/nrn2353.
Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2013. Cerebral processing of voice gender studied using a continuous carryover FMRI design. Cereb. Cortex 23, 958–966. https://doi.org/10.1093/cercor/bhs090.
Cusack, R., Brett, M., Osswald, K., 2003. An evaluation of the use of magnetic field maps to undistort echo-planar images. Neuroimage 18, 127–142. https://doi.org/10.1006/nimg.2002.1281.
Daube, C., Ince, R.A.A., Gross, J., 2019. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol. 29, 1924–1937. https://doi.org/10.1016/j.cub.2019.04.067 e9.
DiMattina, C., Wang, X., 2006. Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. J. Neurophysiol. 95, 1244–1262. https://doi.org/10.1152/jn.00818.2005.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in openSMILE, the munich open-source multimedia feature extractor. MM 2013 - Proc. 2013 ACM Multimed. Conf. 835–838. https://doi.org/10.1145/2502081.2502224.
Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., Andre, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P., 2016. The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7, 190–202. https://doi.org/10.1109/TAFFC.2015.2457417.
Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322 (80-), 970–973. https://doi.org/10.1126/science.1164318.
Gentner, T.Q., Margoliash, D., 2003. Neuronal populations and single cells representing learned auditory objects. Nature 424, 669–674. https://doi.org/10.1038/nature01731.
Ghazanfar, A.A., Turesson, H.K., Maier, J.X., van Dinther, R., Patterson, R.D., Logothetis, N.K., 2007. Vocal-tract resonances as indexical cues in Rhesus monkeys. Curr. Biol. 17, 425–430. https://doi.org/10.1016/j.cub.2007.01.029.
Gil-Da-Costa, R., Martin, A., Lopes, M.A., Mũoz, M., Fritz, J.B., Braun, A.R., 2006. Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. Nat. Neurosci. 9, 1064–1070. https://doi.org/10.1038/nn1741.
Griffiths, T.D., Warren, J.D., 2004. What is an auditory object? Nat. Rev. Neurosci. 5, 887–892. https://doi.org/10.1038/nrn1538.
Hausfeld, L., Riecke, L., Valente, G., Formisano, E., 2018. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. Neuroimage 181, 617–626. https://doi.org/10.1016/j.neuroimage.2018.07.052.
Hebart, M.N., Görgen, K., Haynes, J.D., 2015. The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinform. 8 https://doi.org/10.3389/fninf.2014.00088.
Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N.K., Uludag, K., Kayser, C., 2013. Spatial representations of temporal and spectral sound cues in human auditory cortex. Cortex 49, 2822–2833. https://doi.org/10.1016/j.cortex.2013.04.003.
Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., Chang, E.F., 2016. Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. J. Neurosci. 36, 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016.
Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fMRI: a quantitative evaluation. Neuroimage 16, 217–240. https://doi.org/10.1006/nimg.2001.1054.
Isik, L., Koldewyn, K., Beeler, D., Kanwisher, N., 2017. Perceiving social interactions in the posterior superior temporal sulcus. Proc. Natl. Acad. Sci. U. S. A. 114, E9145–E9152. https://doi.org/10.1073/pnas.1714471114.
Kasper, L., Bollmann, S., Diaconescu, A.O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T. U., Sebold, M., Manjaly, Z.M., Pruessmann, K.P., Stephan, K.E., 2017. The PhysIO toolbox for modeling physiological noise in fMRI data. J. Neurosci. Methods 276, 56–72. https://doi.org/10.1016/j.jneumeth.2016.10.019.
Kriegstein, K.V., Giraud, A.L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. Neuroimage 22, 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020.
Leaver, A.M., Rauschecker, J.P., 2016. Functional topography of human auditory cortex. J. Neurosci. 36, 1416–1428. https://doi.org/10.1523/JNEUROSCI.0226-15.2016.
McDermott, J.H., Wrobleski, D., Oxenham, A.J., 2011. Recovering sound sources from embedded repetition. Proc. Natl. Acad. Sci. U. S. A. 108, 1188–1193. https://doi.org/10.1073/pnas.1004765108.
Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13, 684–701. https://doi.org/10.1006/nimg.2000.0715.
Morosan, P., Schleicher, A., Amunts, K., Zilles, K., 2005. Multimodal architectonic mapping of human superior temporal gyrus. Anat. Embryol. (Berl). 210, 401–406. https://doi.org/10.1007/s00429-005-0029-1.
Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A., 2012. Deconvolving bold activation in event-related designs for multivoxel pattern classification analyses. Neuroimage 59, 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076.
Mumford, J.A., Poline, J.B., Poldrack, R.A., 2015. Orthogonalization of regressors in fMRI models. PLoS One 10, e0126255. https://doi.org/10.1371/journal.pone.0126255.
Norman-Haignere, S.V., McDermott, J.H., 2018. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol. 16, e2005127 https://doi.org/10.1371/journal.pbio.2005127.
Overath, T., Kumar, S., Von Kriegstein, K., Griffiths, T.D., 2008. Encoding of spectral correlation over time in auditory cortex. J. Neurosci. 28, 13268–13273. https://doi.org/10.1523/JNEUROSCI.4596-08.2008.
Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. 18, 903–911. https://doi.org/10.1038/nn.4021.
Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E.G., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 119, 164–174. https://doi.org/10.1016/j.neuroimage.2015.06.050.
Perrodin, C., Kayser, C., Abel, T.J., Logothetis, N.K., Petkov, C.I., 2015. Who is that? brain networks and mechanisms for identifying individuals. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2015.09.002.
Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374. https://doi.org/10.1038/nn2043.
Rothschild, G., Nelken, I., Mizrahi, A., 2010. Functional organization and population dynamics in the mouse primary auditory cortex. Nat. Neurosci. 13, 353–360. https://doi.org/10.1038/nn.2484.
Sadagopan, S., Temiz-Karayol, N.Z., Voss, H.U., 2015. High-field functional magnetic resonance imaging of vocalization processing in marmosets. Sci. Rep. 5, 10950. https://doi.org/10.1038/srep10950.
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 https://doi.org/10.1371/journal.pcbi.1003412.
Schultz, J., Friston, K.J., O'Doherty, J., Wolpert, D.M., Frith, C.D., 2005. Activation in posterior superior temporal sulcus parallels parameter inducing the percept of animacy. Neuron 45, 625–635. https://doi.org/10.1016/j.neuron.2004.12.052.
Theunissen, F.E., Elie, J.E., 2014. Neural processing of natural sounds. Nat. Rev. Neurosci. 15, 355–366. https://doi.org/10.1038/nrn3731.
Toarmino, C.R., Wong, L., Miller, C.T., 2017. Audience affects decision-making in a marmoset communication network. Biol. Lett. 13 https://doi.org/10.1098/rsbl.2016.0934.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289. https://doi.org/10.1006/nimg.2001.0978.
Van Essen, D.C., Dierker, D.L., 2007. Surface-based and probabilistic atlases of primate cerebral cortex. Neuron 56, 209–225. https://doi.org/10.1016/j.neuron.2007.10.015.
Webster, P.J., Skipper-Kallal, L.M., Frum, C.A., Still, H.N., Ward, B.D., Lewis, J.W., 2017. Divergent human cortical regions for processing distinct acoustic-semantic categories of natural sounds: animal action sounds vs. vocalizations. Front. Neurosci. 10 https://doi.org/10.3389/fnins.2016.00579.
Yildiz, I.B., Mesgarani, N., Deneve, S., 2016. Predictive ensemble decoding of acoustical features explains context-dependent receptive fields. J. Neurosci. 36, 12338–12350. https://doi.org/10.1523/JNEUROSCI.4648-15.2016.
Yovel, G., Belin, P., 2013. A unified coding strategy for processing faces and voices. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2013.04.004.