Reconstructing Voice Identity from Noninvasive Auditory Cortex Recordings

Charly Lamothe[1,2] ✉, Etienne Thoret[1,2,3,4], Régis Trapeau[1], Bruno L. Giordano[1], Julien Sein[1,5], Sylvain Takerkart[1], Stéphane Ayache[2], Thierry Artières[2,6,7] ✉, Pascal Belin[1,7] ✉

[1]La Timone Neuroscience Institute UMR 7289, CNRS, Aix-Marseille University, Marseille, France. [2]Laboratoire d'Informatique et Systèmes UMR 7020, CNRS, Aix-Marseille University, Marseille, France. [3]Perception, Representation, Image, Sound, Music UMR 7061, CNRS, Marseille, France. [4]Institute of Language Communication & the Brain, Marseille. [5]Centre IRM-INT@CERIMED, Marseille, France. [6]École Centrale de Marseille, Marseille, France. [7]These authors jointly supervised this work: Thierry Artières, Pascal Belin. ✉e-mail: charlylmth@gmail.com, thierry.artieres@lis-lab.fr, pascal.belin@univ-amu.fr

Abstract

The cerebral processing of voice information is known to engage, in human as well as non-human primates, "temporal voice areas" (TVAs) that respond preferentially to conspecific vocalizations. However, how voice information is represented by neuronal populations in these areas, particularly speaker identity information, remains poorly understood. Here, we used a deep neural network (DNN) to generate a high-level, small-dimension representational space for voice identity—the 'voice latent space' (VLS)—and examined its linear relation with cerebral activity via encoding, representational similarity, and decoding analyses. We find that the VLS maps onto fMRI measures of cerebral activity in response to tens of thousands of voice stimuli from hundreds of different speaker identities and better accounts for the representational geometry for speaker identity in the TVAs than in A1. Moreover, the VLS allowed TVA-based reconstructions of voice stimuli that preserved essential aspects of speaker identity as assessed by both machine classifiers and human listeners. These results indicate that the DNN-derived VLS provides high-level representations of voice identity information in the TVAs.

Introduction

The human voice carries speech, but is also an "auditory face" that carries much valuable information on the stable physical characteristics of the speaker (hereafter, 'identity-related'; Belin et al., 2004, 2011). The ability of listeners to extract identity-related information from a voice such as gender, age, or unique identity even in brief stimuli plays a crucial role in our social interactions. Yet, its neural bases remain poorly understood compared to those of speech processing. Studies over the past two decades have clearly established via complementary neuroimaging techniques that the cerebral processing of voice information involves a set of temporal voice areas (TVAs) in secondary auditory cortical regions of the human (fMRI: Belin et al., 2000, von Kriegstein et al., 2004, Pernet et al., 2015; EEG, MEG: Charest et al., 2009, Capilla et al., 2013, Barbero et al., 2021; Electrophysiology: Rupp et al., 2022, Zhang et al., 2021) as well as macaque (Petkov et al., 2008; Bodin et al., 2021) brain. The TVAs respond more strongly to sounds of voice – with or without speech (Pernet et al., 2015; Rupp et al., 2022; Trapeau et al., 2022)—and categorize voice apart from other sounds (Bodin et al., 2021) but the nature of the information encoded at these stages of cortical processing, especially with respect to speaker identity-related information, remains largely unknown (Blank et al., 2014; Belin et al., 2018).

In recent years, deep neural networks (DNNs) have emerged as a powerful tool for representing complex visual data, such as images (LeCun et al., 2015), videos (Liu et al., 2020) or audio (Chorowski et al., 2019). In the auditory domain, DNNs have been shown to

provide valuable representations—so-called feature or latent spaces—for modeling the cerebral processing of sound (brain encoding) (speech: Kell et al., 2018; Millet et al., 2022; Tuckute & Feather, 2023; semantic content: Caucheteux et al., 2022; Caucheteux & King, 2022; Caucheteux et al., 2023; Giordano et al., 2023; music: Güçlü et al., 2016), or reconstructing the stimuli heard by a participant (brain decoding) (Akbari et al., 2019). They have not yet been used to explain cerebral representations of identity-related information due in part to the focus on speech information (von Kriegstein et al., 2003).

Here, we addressed this challenge by training a 'Variational autoencoder' (VAE; Kingma et Welling, 2014) DNN to reconstruct voice spectrograms from 182,000 250-ms voice samples from 405 different speaker identities in 8 different languages from the CommonVoice database (Ardila et al., 2020). Brief (250 ms) samples were used to emphasize speaker identity-related information in voice, already available after a few hundred milliseconds (Schweinberger et al., 1997; Lavan, 2023), over linguistic information unfolding over longer periods (word, >350 ms; Mcallister et al., 1994). While 250 ms is admittedly short compared to standards of, e.g., computational speaker identification that typically uses 2-3 s samples, this short duration is sufficient to allow near-perfect gender classification and performance levels well above chance for speaker discrimination. This brief duration allowed the presentation of many more stimuli to our participants in the scanner while preserving acceptable behavioral and classifier performance levels.

State-of-the-art studies have primarily relied on task-optimized neural networks (i.e., DNN trained using supervised learning to classify a category from the input) to study sensory cortex processes (Yamins & DiCarlo, 2016; Schrimpf et al., 2018). They can reach high accuracies in brain encoding (Khaligh-Razavi & Kriegeskorte, 2014; Schrimpf et al., 2018; Han et al., 2019). However, there is increasing evidence that unsupervised learning, such as that used for the VAE, also offers potentially valuable computational models for investigating brain processing (Higgins et al., 2021; Zhuang et al., 2021; Millet et al., 2022; Orhan et al., 2022). Thus, latent spaces derived by a VAE, exploited within encoding, representational similarity, and decoding frameworks, offer a potentially promising tool for investigating the representations of voice stimuli in the secondary auditory cortex (Naselaris et al., 2011). Autoencoders learn to compress stimuli with high dimensionality into a lower-dimensional space that nonetheless allows reconstruction of the original stimuli via an inverse transformation learned by the second part of the network called the decoder.

We trained such a model to learn to compress spectrotemporal representations of voice samples into a *voice latent space* (VLS). In order to test whether VLS accounts well for cerebral activity in response to voice stimuli, we scanned three healthy volunteers using fMRI to measure an indirect index of their cerebral activity across 10+ hours of scanning each in response to ~12,000 of the voice samples, denoted *BrainVoice* in the following (different from the ones used to train the DNN). The small number of participants does not allow for generalization at the general population level as in standard fMRI studies. However, it allows testing for replicability as in comparable studies involving 10+ hours of scanning per participant (VanRullen & Reddy, 2019). Different stimulus sets were used across participants to provide a stringent test of replicability based on subject-level analyses. Stimuli consisted of randomly spliced 250-ms excerpts of speech samples from the CommonVoice database (Ardila et al., 2020) by 119 speakers in 8 languages.

We first asked how the VLS could account for the brain responses to speaker identities (encoding) measured in A1 and the TVAs in comparison with a linear autoencoder's latent space (LIN). This approach was chosen to compare a representation learned linearly under similar conditions (same input data, learning algorithm, reconstruction objective and latent space size) with the VLS, which has non-linear transformations and a regularized latent space. For this, we used a general linear model (GLM) of fMRI responses to the speaker identities, resulting in one voxel activity map per speaker (Figure 2-S1). Then, we computed the average VLS coordinates of the fMRI voice stimuli for each speaker identity, which may be seen as a

speaker representation in the VLS (see *Identity-based and stimulus-based representations* section). Next, we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker's VLS coordinates. As VAE achieves compression through a series of nonlinear transformations (Wetzel, 2017), we choose to contrast its results with a linear autoencoder's latent space. This method has previously been applied to fMRI-based image reconstructions (Cowen et al., 2014; VanRullen & Reddy, 2019; Mozafari et al., 2020).

The extent to which the VLS allows linearly predicting the fMRI recordings does not provide insight into the representational geometries, i.e., the differences between the patterns of cerebral activity for speaker identity. We addressed this question by using representational similarity analysis (RSA; Kriegeskorte et al., 2008) to test which model better accounts for the representational geometry for voice identities in the auditory cortex. Using RSA as a model comparison framework is relevant to examining the brain-model relationship from complementary angles (Diedrichsen & Kriegeskorte, 2017; Giordano et al., 2023; Tuckute & Feather, 2023). We built speaker x speaker representational dissimilarity matrices (RDMs) capturing pairwise differences in cerebral activity or model predictions between all pairs of speakers; then, we examined how well the LIN and VLS-derived RDMs correlated with the cerebral RDMs from A1 and the TVAs.

A robust test of the validity of models of brain activity, and a long-standing goal in computational neurosciences, is the reconstruction of a stimulus presented to a participant from the evoked brain responses. While reconstruction of visual stimuli (images, videos) from cerebral activity has been performed by a number of groups (VanRullen & Reddy, 2019; Mozafari et al., 2020; Le et al., 2022; Gaziv et al., 2022; Dado et al., 2022; Chen et al., 2023), validating the DNN-derived representational spaces, comparable work in the auditory domain is scarce, almost exclusively concentrated on linguistic information (Santoro et al., 2017). Akbari et al. (2019) used a DNN to reconstruct speech stimuli based on ECoG recording of auditory cortex activity, an invasive method compared to techniques like fMRI. They obtained a good phonetic recognition rate but chance-level gender categorization performance from reconstructed spectrograms and no evaluation of speaker identity discrimination.

Here, we built on the linear relationship uncovered in our encoding analysis between the VLS and the fMRI recordings to invert it and try to predict VLS coordinates from the recorded fMRI data; then, using the decoder, we reconstructed the spectrograms of stimuli presented to the participants (Wu et al., 2006; Naselaris et al., 2011). The voice identity information available in the reconstructed stimuli was finally assessed by human listeners using both machine learning classifiers and behavioral tasks (Figure 4).

Results

Voice Information in the Voice Latent Space (VLS). Figure 1a shows the architecture of the VAE, with its encoder that reduces an input spectrogram to a highly compressed, 128-dimension *voice latent space* (VLS) representation and its decoder that reconstructs the spectrogram from this VLS representation. We selected this latent space size as it was the first value that produced satisfactory reconstructions. Points in the VLS correspond to voice samples with different identities and phonetic content. A line segment in the VLS contains points corresponding to perceptual interpolations between its two extremities (Figure 1b; Supplementary Audio 1). VLS coordinates of samples presented to the participants averaged by speaker identity suggest that a major organizational dimension of the latent space is voice gender (Figure 1b) (colored by age or language in Figure 1-S1).

In order to probe the informational content of the VLS, linear classifiers were trained to categorize the voice stimuli from 405 speakers by gender (2 classes), age (2 classes: before/above 30 years old, the median age in our sample) or identity (119 classes, cf Methods) based on VLS coordinates, or their LIN features (Figure 1c,d,e; we aggregated the stimuli from the 3 participants; for each model, we computed the latent space of each stimulus and

averaged the latent spaces by speaker identity, leading to 405 128-dimensional vectors. We then trained linear classifiers using a 5-fold cross-validation scheme, see *Characterization of the autoencoder latent space*). The decoding accuracy was significantly above chance level (Wilcoxon signed-rank test, all W=15, p=0.03125) for all classifications (LIN: gender (mean accuracy ± s.d.) = 97.64±1.77%; age: 64.39±4.54%; identity: 40.52±9.14%; VLS: gender: 98.59±1.19%; age: 67.31±4.86%; identity: 38.40±8.75%).

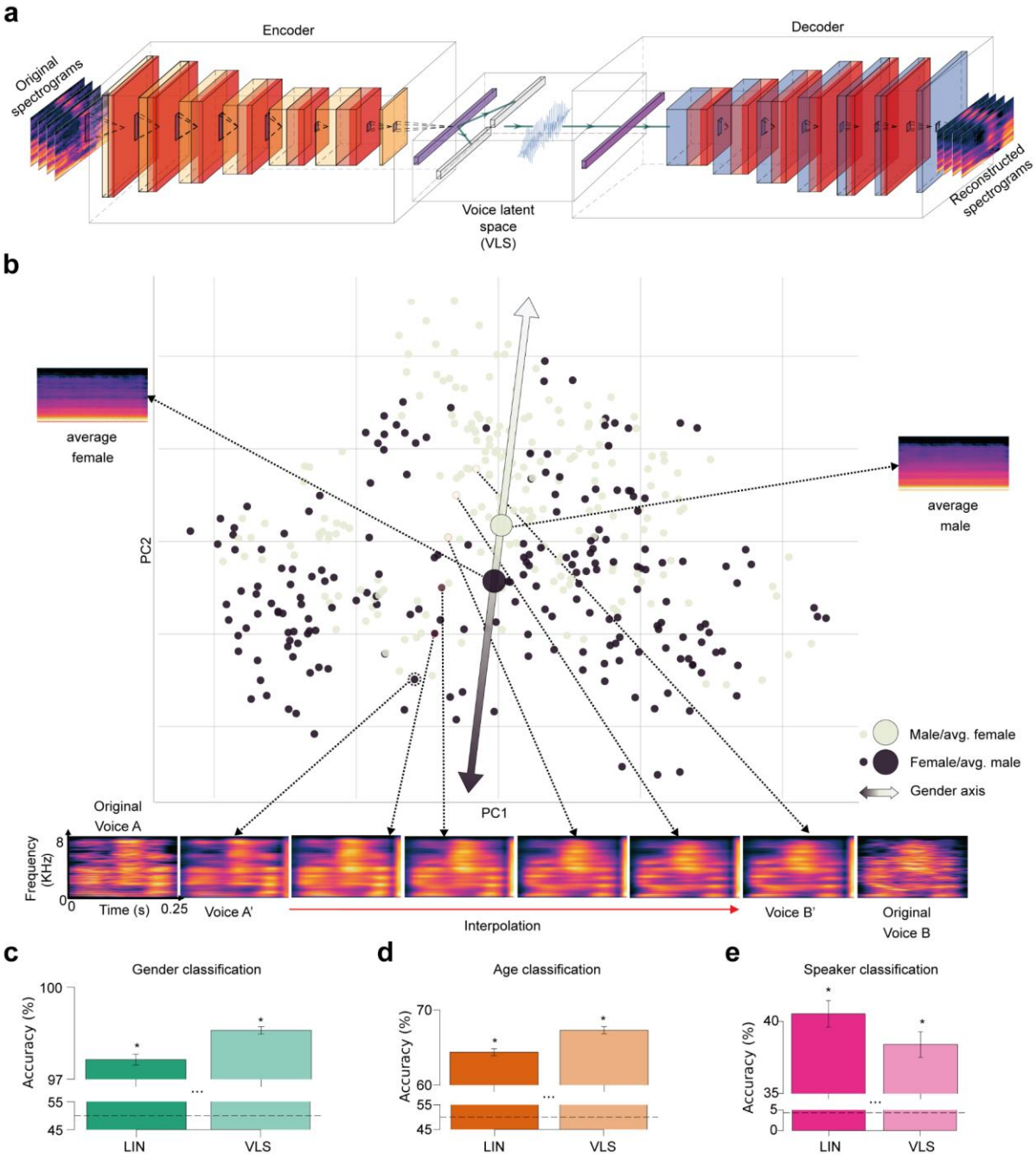

Figure 1. DNN-derived Voice Latent Space (VLS). a, Variational autoencoder (VAE) Architecture. Two networks learned complementary tasks. An encoder was trained using 182K voice samples to compress their spectrogram into a 128-dimension representation, the voice latent space (VLS), while a decoder learned the reverse mapping. The network was trained end-to-end by minimizing the difference between the original and reconstructed spectrograms. **b,** Distribution of the 405 speaker identities along the first 2 principal components of the VLS coordinates from all sounds, averaged by speaker identity. Each disk represents a speaker's

identity colored by gender. PC2 largely maps onto voice gender (ANOVAs on the first two components: PC1: F(1, 405)=0.10, p=.74; PC2: F(1, 405)=11.00, p<.001). Large disks represent the average of all male (black) or female (gray) speaker coordinates, with their associated reconstructed spectrograms (note the flat fundamental frequency (f_0) and formant frequencies contours caused by averaging). The bottom of the spectrograms illustrates an interpolation between stimuli of two different speaker identities: spectrograms at the extremes correspond to two original stimuli (A, B) and their VLS-reconstructed spectrograms (A', B'). Intermediary spectrograms were reconstructed from linearly interpolated coordinates between those two points in the VLS (red line) (cf. Supplementary Audio 1). **c,d e,** Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (young/adult, chance level: 50%), or identity (119 identities, chance level: 0.84%) based on VLS or Linear model (LIN) coordinates. Error bars indicate the standard error of the mean (s.e.m) across 5 folds. All ps<0.05. The horizontal black dashed lines indicate chance levels. *: p<0.05.

Thus, despite its low number of dimensions (each input spectrogram has 401x21=8421 parameters and is summarized in the VLS by a mere 128 dimensions), the VLS appears to meaningfully represent the different sources of voice information perceptually available in the vocal stimuli. This representational space, therefore, constitutes a relevant candidate for linearly modeling voice stimulus representations by the brain.

Brain Encoding. For assessing generalization performances of decoding models and brain-based reconstruction, six test stimuli were repeated more often (60 times) for each participant to provide robust estimates of their induced cerebral activity (see Methods). We first modeled these responses to voice using a general linear model (GLM) (Friston et al., 1994) with several nuisance regressors as an initial denoising step (Figure 2-S2), then used a second GLM modeling cerebral responses to the different speaker identities (Figure 2-S1a), resulting in one voxel activity map per speaker (Figure 2-S1b). We independently localized in each participant several regions of interest (ROIs) on which subsequent analyses were focused: the anterior, middle and posterior TVAs in each hemisphere (individually localized via an independent 'voice localizer scan' and MNI coordinates provided in Pernet et al., 2015; Figure 2-S1c) as well as primary auditory cortex (A1) (using a probabilistic map in MNI space (Penhune et al., 1996; Figure 1-S2d).

We used a linear voxel-based encoding model to test whether VLS linearly maps onto cerebral responses to speaker identities measured with fMRI in the different ROIs. A regularized linear regression model (cf. Methods) was trained on a subset of the data (5-fold cross-validation scheme) to predict the voxel maps for each speaker identity. For each fold, the trained model was tested on the held-out speaker identities (Figure 2a). The model's performance was assessed for each ROI using the Pearson correlation score between each voxel's actual and predicted responses (Schrimpf et al., 2021). Similar predictions were tested with features derived from LIN, as well as with more recent unsupervised/self-supervised speech models that have been shown to align to auditory responses to speech in the temporal cortex: Wav2Vec (Millet et al., 2022) and HuBERT (Li et al., 2023) (Figure 2-S3). Figure 2b shows the distribution of correlation coefficients obtained for each of the ROIs for the 2 sets of features across voxels, hemispheres, and participants.

One-sample t-tests showed that the means of Fisher z-transformed coefficients for both LIN features and VLS were significantly higher than zero (LIN: A1 t(197)=7.25, p<.0001, pTVA t(175)=4.49, p<.0001, mTVA t(164)=9.12, p<.0001 and aTVA t(147)=6.81, p<.0001; VLS: A1 t(197)=4.76, p<.0001, mTVA t(164)=10.12, p<.0001 and aTVA t(147)=5.52, p<.0001 but not pTVA t(175)=-1.60) (Supplementary Tables 2-3).

A mixed ANOVA performed on the Fisher z-transformed coefficients with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors showed a significant effect of Feature (F(3, 683)=56.65, p<.0001), a significant effect of ROI (F(3, 683)=18.50, p<.0001), and a moderate interaction Feature x ROI (F(3, 683)=5.25, p<.01). Post-hoc comparisons revealed that the

mean of correlation coefficients was higher for LIN than for VLS in A1 (t(197)=4.02, p<.0001), pTVA (t(175)=6.64, p<.0001), aTVA (t(147)=3.78, p<.001) but not in mTVA (t(164)=0.58) (Supplementary Table 4); and that the voxel patterns are better predicted in mTVA than in A1 for both models (LIN: t(361)=2.36, p<.05); VLS: t(361)=4.91, p<.0001) (Supplementary Table 5). However, inspecting the distribution of model-voxel correlations, we found that both models account for different parts of the voice identity responses and differ across ROIs (Figure 2c).

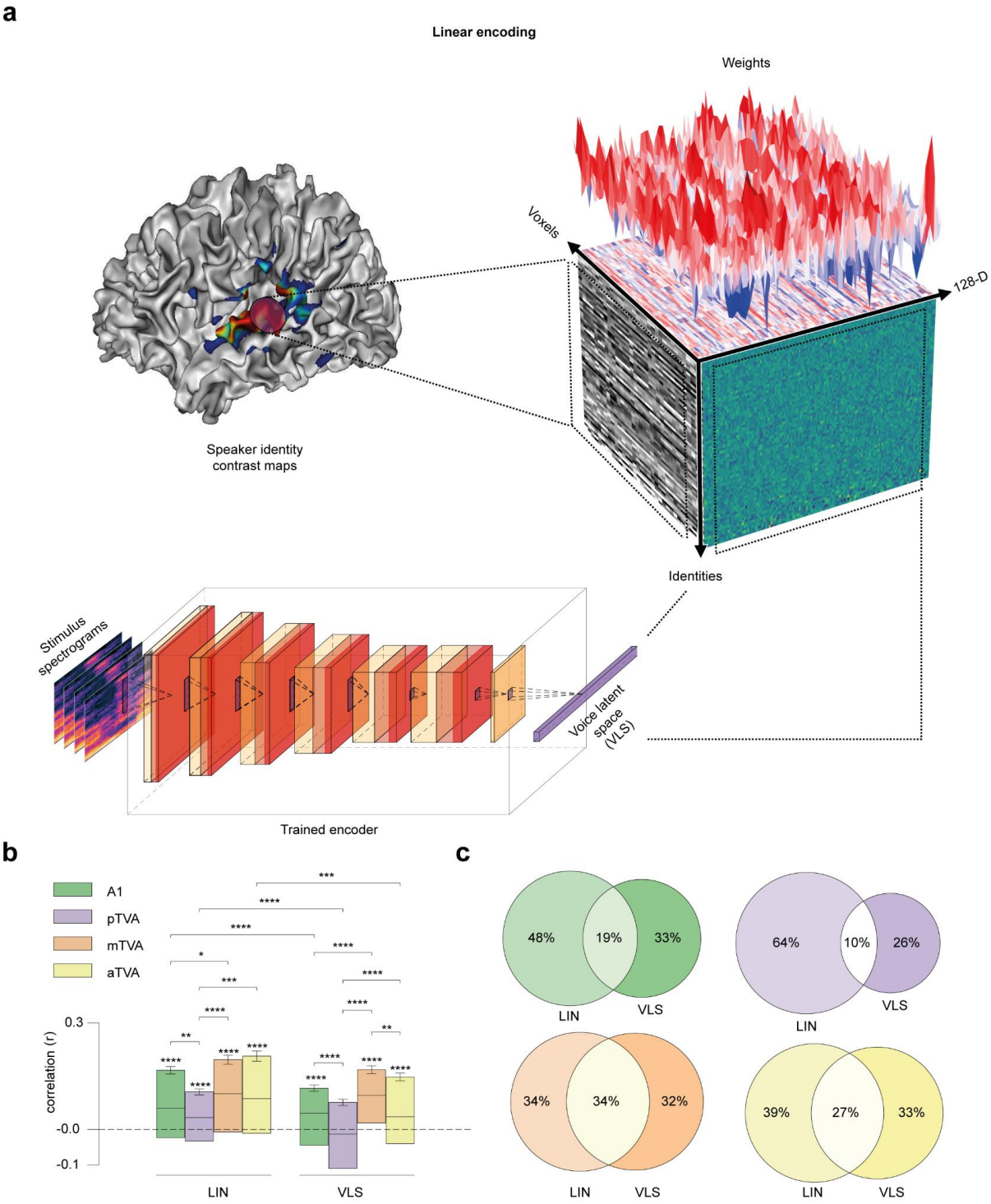

Figure 2. Predicting brain activity from the VLS. a, Linear brain activity prediction from VLS for ~135 speaker identities in the different ROIs. We first fit a GLM to predict the Blood Oxygenation Level-Dependant (BOLD) responses to each voice speaker identity. Then, using the trained encoder, we computed the average VLS coordinates of the voice stimuli presented

to the participants based on speaker identity. Finally, we trained a linear voxel-based encoding model to predict the speaker voxel activity maps from the speaker VLS coordinates. The cube illustrates the linear relationship between the fMRI responses to speaker identity and the VLS coordinates. The left face of the cube represents the activity of the voxels for each speaker's identity, with each line corresponding to one speaker. The right face displays the VLS coordinates for each speaker's identity. The cube's top face shows the encoding model's weight vectors. **b,** Encoding results. For each region of interest, the model's performance was assessed using the Pearson correlation score between the true and the predicted responses of each voxel on the held-out speaker identities. Pearson's correlation coefficients were computed for each voxel on the speakers' axis and then averaged across hemispheres and participants. Similar predictions were tested with the LIN features. Error bars indicate the standard error of the mean (s.e.m) across voxels. *$p < 0.05$; **$p < 0.01$; **$p < 0.001$; ****$p < 0.0001$. **c,** Venn diagrams of the number of voxels in each ROI with the LIN, the VLS, or both models. For each ROI and each voxel, we checked whether the test correlation was higher than the median of all participant correlations (intersection circle), and if not, which model (LIN or VLS) yielded the highest correlation (left or right circles).

Representational Similarity Analysis (RSA). For RSA, we built speaker x speaker representational dissimilarity matrices (RDMs), capturing for each ROI the dissimilarity in voxel space between each pair of speaker voxel maps ('brain RDMs'; cf. Methods) using Pearson's correlation (Walther et al., 2016). We compared these four bilateral brain RDMs (A1, aTVA, mTVA, pTVA) to two 'model RDMs' capturing speaker pairwise feature differences predicted by LIN and the VLS (Figure 3a) built using cosine distance (Xing et al., 2015; Bhattacharya et al., 2017; Wang et al., 2018). Figure 3b shows for each ROI the Spearman correlation coefficients between the brain RDMs and the two model RDMs for each participant and hemisphere (Kriegeskorte et al., 2008; Figure 3c for an example of brain-model correlation).

To correct for multiple comparisons across models, these brain-model correlation coefficients were compared to zero using a 'maximum statistics' approach based on random permutations of the model RDMs' rows and columns (Maris & Oostenveld, 2007; cf. Methods; Figure 3b). For the LIN model, only one brain-model RDM correlation was significantly different from zero (one-tailed test, corrected for multiple comparisons across models): in mTVA, right hemisphere in S3 ($p=.0500$). For the VLS model, in contrast, 5 significant brain-model RDM correlations were observed in all four ROIs: in A1, right hemisphere in S3 ($p=.0142$); pTVA: right hemisphere in S3 ($p=.0160$); mTVA: left hemisphere in S3 ($p=.007$); aTVA: left hemispheres in S1 ($p=.0417$) and S3 ($p=.0001$) (Supplementary Table 6).

A two-way repeated-measures ANOVA with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on the Fisher z-transformed correlation coefficients showed a tendency towards a significant effect of Feature ($F(1, 2)=22.53$, $p=.04$), and no ROI ($F(3, 6)=1.79$, $p=.30$) or interaction effects ($F(3, 6)=1.94$, $p=.22$). We compared the correlation coefficients between the VLS and LIN models within participants and hemispheres using one-tailed tests, based on the a priori hypothesis that the VLS models would exhibit greater brain-model correlations than the LIN models (cf. Methods). The results revealed two significant differences in one of the three participants, both favoring the VLS model (S3: right pTVA, $p=.0366$; left aTVA, $p=.00175$) (Supplementary Table 7).

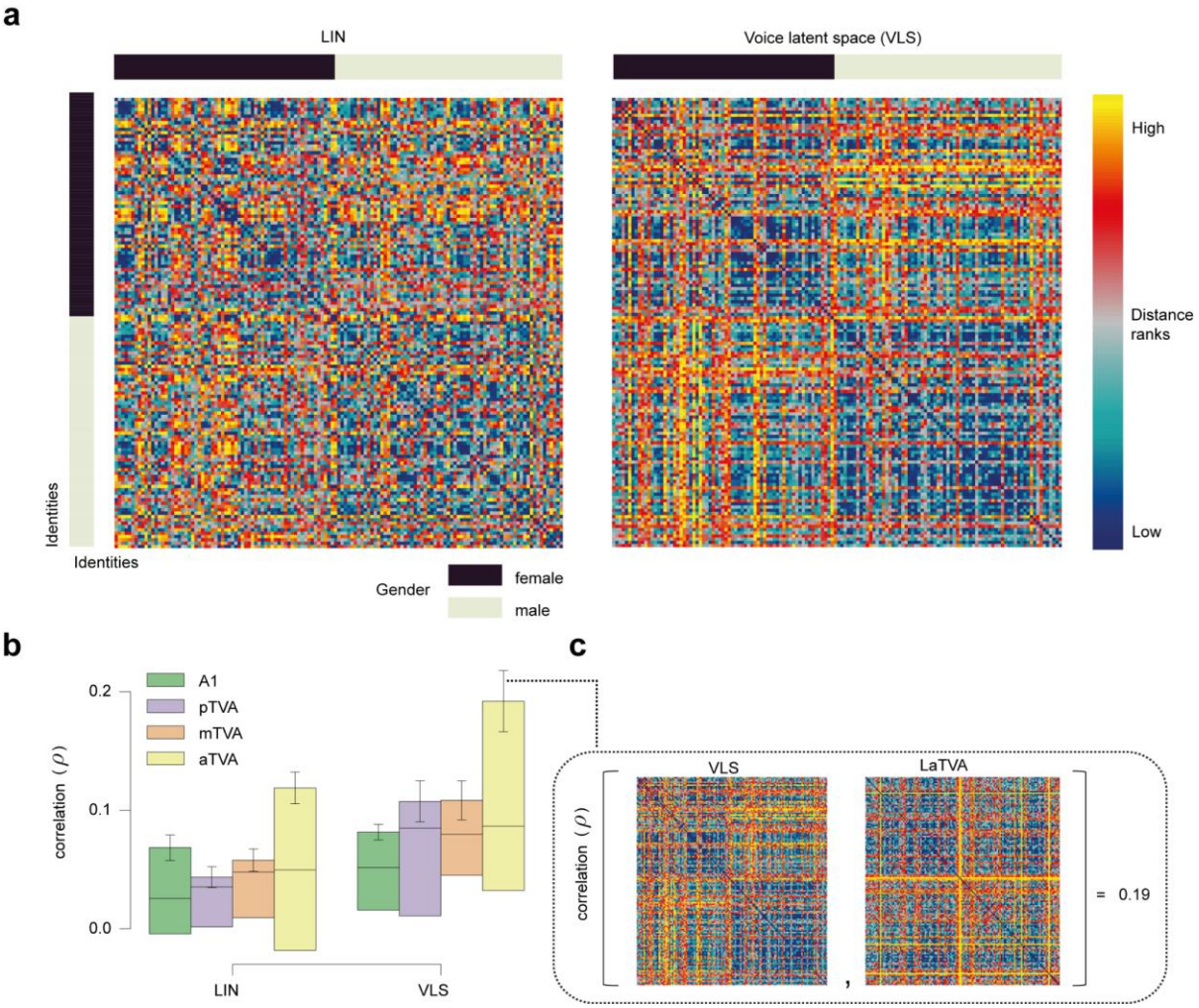

Figure 3. The VLS better explains representational geometry for voice identities in the TVAs than the linear model. a, Representational dissimilarity matrices (RDMs) of pairwise speaker dissimilarities for ~135 identities (arranged by gender, cf. sidebars), according to LIN and VLS. **b,** Spearman correlation coefficients between the brain RDMs for A1, the 3 TVAs, and the 2 model RDMs. Error bars indicate the standard error of the mean (s.e.m) across brain-model correlations. **c,** Example of brain-model RDM correlation in the TVAs. The VLS RDM and the brain RDM yielding one of the highest correlations (LaTVA) are shown in the insert.

Decoding and Reconstruction. We finally inverted the brain-VLS relationship to predict linearly VLS coordinates based on fMRI measurements (Figure 4a; see 'Brain decoding' in Methods) and reconstructed via the trained decoder the spectrograms of 18 Test Stimuli (3 participants x 6 stimuli per participant; see Figure 4b, and Supplementary Audio 2; audio estimated from spectrogram through phase reconstruction).

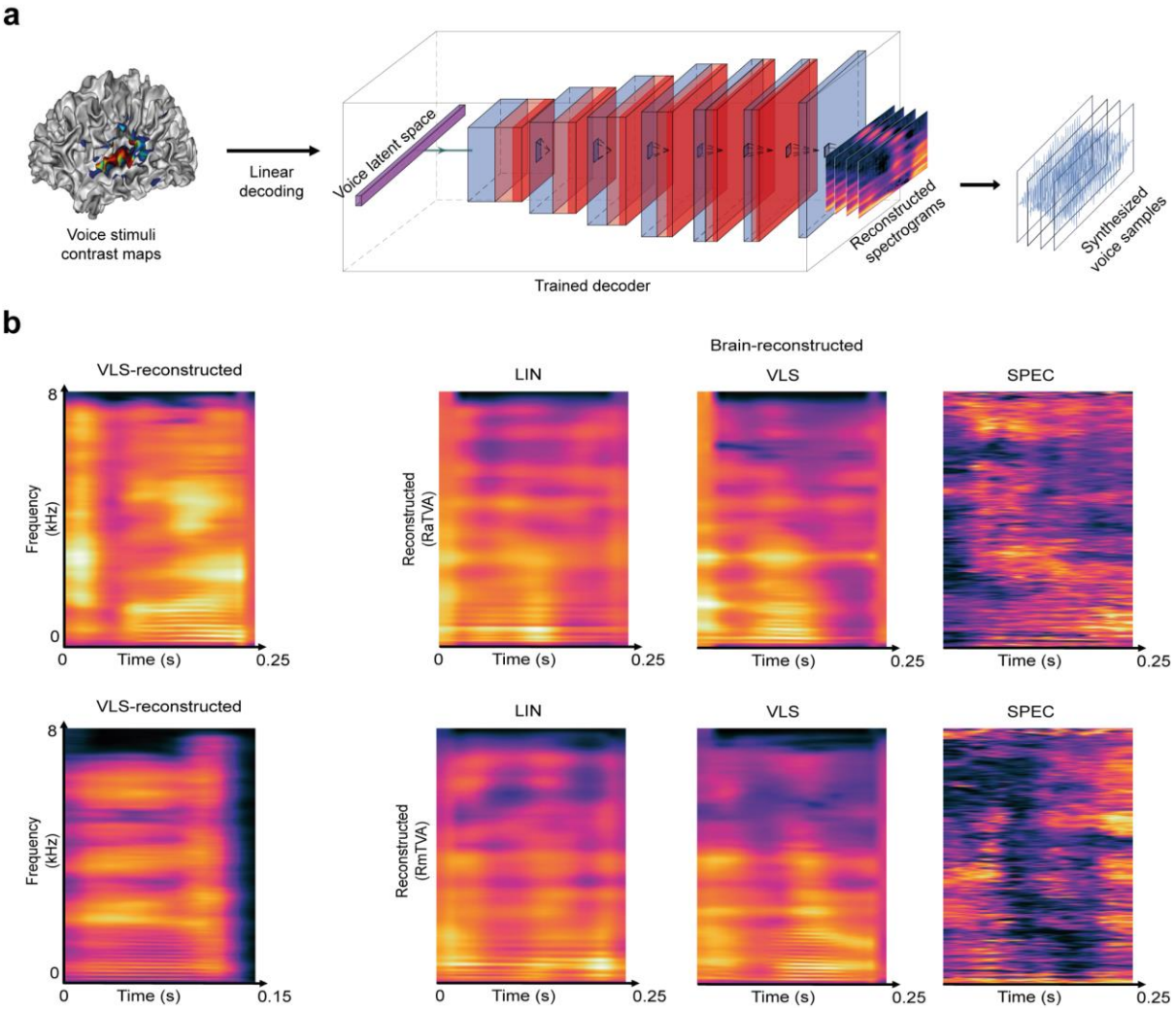

Figure 4. Reconstructing voice identity from brain recordings. a, A linear voxel-based decoding model was used to predict the VLS coordinates of 18 Test Stimuli based on fMRI responses to ~12,000 Train stimuli in the different ROIs. To reconstruct the audio stimuli from the brain recordings, the predicted VLS coordinates were then fed to the trained decoder to yield reconstructed spectrograms, synthesized into sound waveforms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983). **b,** Reconstructed spectrograms of the stimuli presented to the participants. The left panels show the spectrogram of example original stimuli reconstructed from the VLS, and the right panels show brain-reconstructed spectrograms via LIN or VLS autoencoder-based representations, and SPEC, direct regression from the audio spectrograms (cf. Supplementary Audio 2).

We first assessed the nature of the reconstructed stimuli by using a DNN trained to categorize natural audio events (Howard et al., 2017): all reconstructed versions of the 18 Test Stimuli were categorized as 'speech' (1 class out of 521 - no ‘voice’ classes). To evaluate the preservation of voice identity information in the reconstructed voices, pre-trained linear classifiers were used to classify the speaker gender (2 classes), age (2 classes), and identity (17 classes; one identity was shared across participants) of the 18 reconstructed Test Stimuli. The mean of the accuracy distribution obtained across random classifier initializations (20 per ROI) used on the stimuli reconstructed from the induced brain activity was significantly above chance level for gender (LIN: pTVA (mean accuracy ± s.d.): 72.08±5.48, t(39)=25.15; VLS: A1: 61.11±2.15, t(39)=32.25; pTVA: 63.89±2.78, t(39)=31.22), age (LIN: pTVA: 54.58±4.14, t(39)=6.90; aTVA: 63.96±12.55, t(39)=6.94; VLS: pTVA: 65.00±7.26, t(39)=12.89; aTVA: 60.42±5.19, t(39)=12.54) and identity (LIN: A1: 9.20±9.23, t(39)=2.24; pTVA: 9.48±4.90,

t(39)=4.59; aTVA: 9.41±6.28, t(39)=3.51; VLS: pTVA: 16.18±7.05, t(39)=9.11; aTVA: 8.23±4.70, t(39)=3.12) (Figure 5a-c; Supplementary Tables 8-10).

Two-way ANOVAs with Feature (VLS, LIN) and ROI (A1, pTVA, mTVA, aTVA) as factors performed on classification accuracy scores (gender, age, identity) revealed for gender classifications significant effects of Feature F(1, 312)=12.82, p<.0005) and ROI (gender: F(3, 312)=245.06, p<.0001; age: F(3, 312)=64.49, p<.0001; identity: F(3, 312)=14.49, p<.0001), as well as Feature x ROI interactions (gender: F(3, 312)=56.74, p<.0001; age: F(3, 312)=4.31, p<.001; identity: F(3, 312)=8.82, p<.0001). Post-hoc paired t-tests indicated that the VLS was better than LIN in preserving gender, age and identity information in at least one TVA compared with A1 (gender: aTVA: t(39)=5.13, p<.0001; age: pTVA: t(39)=9.78, p<.0001; identity: pTVA: t(39)=4.01, p<.0005) (all tests in Supplementary Table 11). Post-hoc two sample t-tests comparing ROIs revealed significant differences in all classifications, in particular with pTVA outperforming other ROIs in gender (LIN: pTVA vs A1: t(78)=22.40, p<.0001; pTVA vs mTVA: t(78)=10.92, p<.0001; pTVA vs aTVA: t(78)=31.47, p<.0001; VLS: pTVA vs A1: t(78)=4.94, p<.0001; pTVA vs mTVA: t(78)=13.96, p<.0001; pTVA vs aTVA: t(78)=22.06, p<.0001), age (LIN: pTVA vs A1: t(78)=7.26, p<.0001; pTVA vs mTVA: t(78)=10.11, p<.0001; VLS: pTVA vs A1: t(78)=5.71, p<.0001; pTVA vs mTVA: t(78)=10.11, p<.0001; pTVA vs aTVA: t(78)=3.21, p<.005) and identity (LIN: pTVA vs mTVA: t(78)=2.27, p<.05; VLS: pTVA vs A1: t(78)=6.45, p<.0001; pTVA vs mTVA: t(78)=6.62, p<.0001; pTVA vs aTVA: t(78)=5.85, p<.0001) (Supplementary Table 12).

We further evaluated voice identity information in the reconstructed stimuli by testing human participants (n=13) in a series of 4 online experiments assessing the reconstructed stimuli on (i) naturalness judgment, (ii) gender categorization, (iii) age categorization, and (iv) speaker categorization (cf. Methods). The naturalness rating task showed that the VLS-reconstructed stimuli sounded more natural compared to LIN-reconstructed ones, as revealed by a two-way repeated-measures ANOVA (factors: Feature and ROI) with a strong effect of Feature (F(1, 12)=53.72, p<.0001) and a small ROI x Feature interaction (F(3, 36)=5.36, p<.005). Post-hoc paired t-tests confirmed the greater naturalness of VLS-reconstructed stimuli in both A1 and the TVAs (all ps<.0001) (Figure 5g). For the gender task, one-sample t-tests showed that categorization of the reconstructed stimuli was only significantly above chance level for the VLS (A1: (mean accuracy ± s.d.) 55.77±10.84, t(25)=2.66, p<.01; pTVA: 61.75±7.11, t(25)=8.26, p<.0001; aTVA: 55.13±9.23, t(25)=2.78, p<.01). Regarding the age and speaker categorizations, results also indicated that both the LIN- and VLS-reconstructed stimuli yielded above-chance performance in the TVAs (age: LIN: aTVA, 55.77±14.95, t(25)=1.93, p<.05; VLS: aTVA, 63.14±11.82, t(25)=5.56, p<.0001; identity: LIN: pTVA: 54.38±9.34, t(17)=1.93, p<.05; VLS: pTVA: 63.33±6.75, t(17)=8.14, p<.0001) (Supplementary Tables 13-15). Two-way repeated-measures ANOVAs revealed a significant effect of ROI for all categories (gender: F(3, 27)=5.90, p<.05; age: F(3, 36)=14.25, p<.0001; identity: F(3, 24)=38.85, p<.0001), and a Feature effect for gender (F(1, 9)=43.61, p<.0001) and identity (F(1, 8)=14.07, p<.001), but not for age (F(1, 12)=4.01, p=0.07), as well as a ROI x Feature interaction for identity discrimination (F(3, 24)=3.52, p<.05) (Supplementary Tables 16-17 for the model and ROI comparisons).

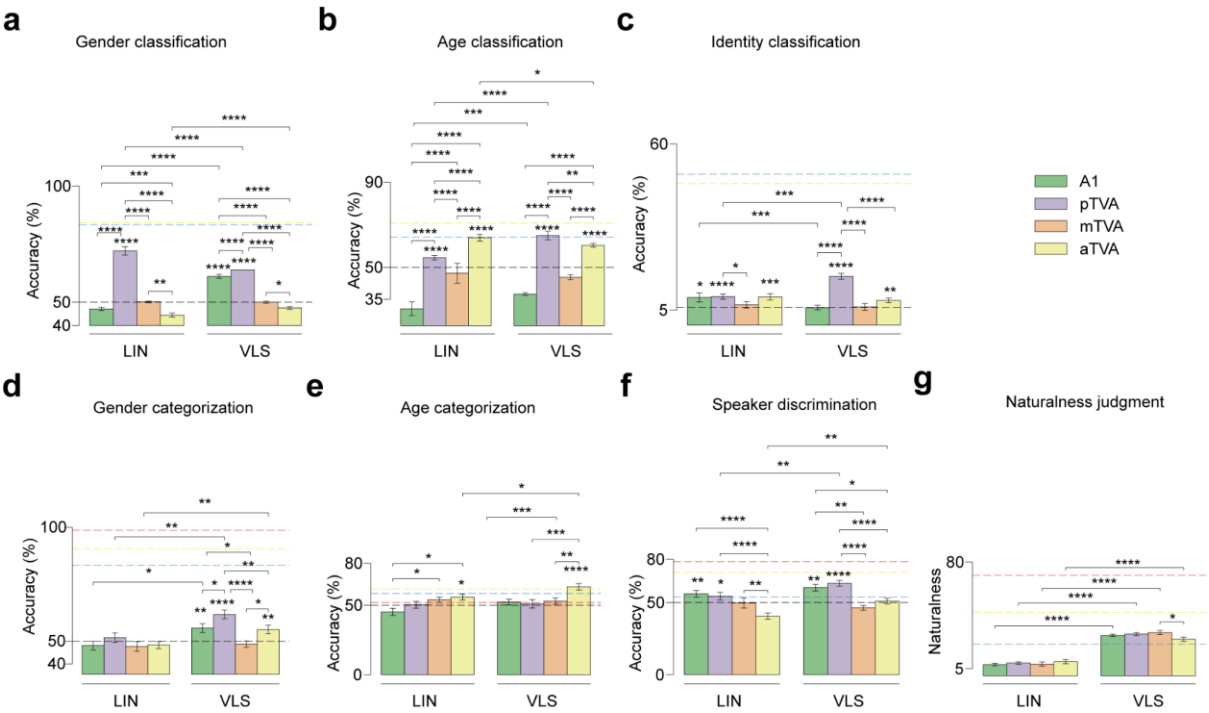

Figure 5. Behavioural and machine classification of the reconstructed stimuli. a,b,c, Decoding voice identity information in brain-reconstructed spectrograms. Performance of linear classifiers at categorizing speaker gender (chance level: 50%), age (chance level: 50%), and identity (17 identities, chance level: 5.88%). Error bars indicate s.e.m across 40 random classifier initializations per ROI (instance of classifiers; 2 hemispheres x 20 seeds). The horizontal black dashed line indicates the chance level. The blue and yellow dashed lines indicate the LIN and VLS ceiling levels, respectively. *p < .05; **p < .001, ***p < .001; ****p < .0001. **d,e,f,** Listener performance at categorizing speaker gender (chance level: 50%) and age (chance level: 50%), and at identity discrimination (2 forced choice task, chance level: 50%) in the brain-reconstructed stimuli. Error bars indicate s.e.m across participant scores. The horizontal black dashed line indicates the chance level, while the red, blue, and yellow dashed lines indicate the ceiling levels for the original stimuli, the LIN-reconstructed and the VLS-reconstructed, respectively. *p < .05; **p < .01; ***p < .001, ***p < .0001. **g,** Perceptual ratings of voice naturalness in the brain-reconstructed stimuli' as assessed by human listeners, between 0 and 100 (zoomed between 5-80). *p < .05, ****p < .0001.

Discussion

In this study we examined to what extent the cerebral activity elicited by brief voice stimuli can be explained by machine-learned representational spaces, specifically focusing on identity-related information. We trained a linear model and a DNN model to reconstruct 100,000s of short voice samples from 100+ speakers, providing low-dimensional spaces (LIN and VLS), which we related to fMRI measures of cerebral response to thousands of experimental stimuli. We find: (i) that 128 dimensions are sufficient to explain a portion of the brain activity elicited by the voice samples and yield brain-based voice reconstructions that preserve identity-related information; (ii) that the DNN-derived VLS shows improve performance relative to the LIN space, particularly in yielding more brain-like representational spaces and more naturalistic voice reconstructions; (iii) that different ROIs have different degrees of brain-model relationship, with marked differences between A1 and the a, m, and pTVAs.

Low-dimensional spaces generated by machine learning have been used to approximate cerebral face representations and reconstruct recognizable faces based on fMRI (VanRullen & Reddy, 2019; Dado et al., 2022). In the auditory domain, however, they have mainly been

used with a focus on linguistic (speech) information, ignoring identity-related information (but see Akbari et al., 2019). Here, we applied them to brief voice stimuli–with minimal linguistic content but already rich identity-related information–and found that as little as 128 dimensions account reasonably well for the complexity of cerebral responses to thousands of these voice samples as measured by fMRI (Figure 2). LIN and VLS both showed brain-like representational geometries, particularly the VLS in the aTVAs (Figure 3). They made possible what is, to our knowledge, the first fMRI-based voice reconstructions to preserve voice-related identity information such as gender, age, or even individual identity, as indicated by above-chance categorization or discrimination performance by both machine classifiers (Figure 5a-c) and human listeners (Figure 5d-f).

Estimation of fMRI responses (encoding) by LIN yielded correlations largely comparable to those by VLS (Figure 2b), although many voxels were only explained by one or the other space (Figure 2c). However, in the RSA, VLS yielded higher overall correlations with brain RDMs (Figure 3), suggesting a representational geometry closer to that instantiated in the brain than LIN. Further, VLS-reconstructed stimuli sounded more natural than the LIN-reconstructed ones (Figure 5g) and yielded both the best speaker discrimination by listeners (Figure 5f) and speaker classification by machine classifiers (Figure 5c). Unlike LIN, which was generated via linear transforms, VLS was obtained through a series of nonlinear transformations (Wetzel, 2017). The fact that the VLS outperforms LIN in decoding performance indicates that nonlinear transformation is required to better account for the brain representation of voices (Naselaris et al., 2011; Cowen et al., 2014; Han et al., 2019).

Comparisons between ROIs revealed important differences between A1 and the a, m, and pTVAs. For both LIN and VLS, fMRI signal (encoding) predictions were more accurate for the mTVAs than for A1, and for A1 than for the pTVAs (Figure 2b). The aTVAs yielded the highest correlations with the models in the RSA (Figure 3). Stimulus reconstructions (Figure 4) based on the TVAs also yielded better gender, age, and identity classification than those based on A1, with gender and identity best preserved in the pTVA-, and to a lesser extent, in the aTVA-based reconstructions (Figure 5). These results show that the a and pTVAs not only respond more strongly to vocal sounds than A1, but they also represent identity-related information in voice better than mTVA, which was previously anticipated in some neuroimaging studies (Gender: Charest et al., 2013; Identity: Belin & Zatorre, 2003; Maguinness et al., 2018; Roswandowitz et al., 2018; Aglieri et al., 2021). Moreover, several recent studies, using intracranial recordings, either through ECoG electrode grids (Zhang et al., 2021) or sEEG recordings (Rupp et al., 2022), found evidence that supports the idea of a hierarchical organization of voice patches in the temporal lobe, where the information flow starts from the mTVA patches and moves in two directions: one from mTVA to the anterior TVA (aTVA) and the other one from mTVA to posterior TVA (pTVA).

Overall, we show that a DNN-derived representational space provides an interesting approximation of the cerebral representations of brief voice stimuli that can preserve identity-related information. We find it remarkable that such results could be obtained to explain sound representations despite the poor temporal resolution of fMRI. Future work combining more complex architectures to time-resolved measures of cerebral activity, such as magneto-encephalography (Défossez et al., 2023) or ECoG (Pasley et al., 2012), will likely yield better models of the cerebral representations of voice information.

Methods

Experimental procedure overview

Three participants attended 13 MRI sessions each. The first session was dedicated to acquire high-resolution structural data, as well as to identify the voice-selective areas of each participant using a 'voice localizer' based on different stimuli than those in the same experiment (Pernet et al., 2015; see below).

The next 12 sessions began with the acquisition of two fast structural scans for inter-session realignment purposes, followed by six functional runs, during which the main stimulus set of the experiment was presented. Each functional run lasted approximately 12 minutes, during which 240 experimental stimuli were presented in a rapid event-related design with a jittered inter-stimulus interval of 2.8-3.2 s.

Participants 1 and 2 attended all scanning sessions (72 functional runs in total); due to unforeseen personal health reasons, Participant 3's participation was limited to 24 runs.

Participants were instructed to stay still in the scanner while listening to the stimuli. To maintain participants' awareness during functional scanning, they were asked to press an MRI-compatible button each time they heard the same stimulus two times in a row, a rare event occurring 3% of the time (correct button hits (median accuracy ± s.d.): S1=96.67±7.10, S2=100.00±0.89, S3=95.00±3.68).

Scanning sessions were spaced apart by at least two days to avoid possible auditory fatigue due to the exposure to scanner noise. To ensure that participants' hearing abilities did not vary across scanning sessions, hearing thresholds were measured before each session using a standard audiometric procedure (Martin & Champlin, 2000; ISO 2004) and compared with the thresholds obtained prior to the first session.

Participants

This study was part of the project 'Réseaux du Langage' and was promoted by the National Center for Scientific Research (CNRS). It has been given favorable approval by the local ethics committee (Comité de Protection des Personnes Sud-Méditerranée) on the date of 13th February 2019. The National Agency for Medicines (ANSM) has been informed of this study, which is registered under the number 2017-A03614-49. Three native French human speakers were scanned (all females, 26-33 years old). Participants gave written informed consent and received a compensation of 40€ per hour for their participation. All were right-handed and no one had a hearing disorder or neurological disease. All participants had normal hearing thresholds below 15 dB HL, for octave frequencies between 0.125 and 8 kHz.

Stimuli

The auditory stimuli were divided into two sequences. One 'voice localizer' sequence to identify the voice-selective areas of each participant (Pernet et al., 2015) and 'main voice stimuli'.

Voice localizer stimuli. The voice localizer stimuli consisted of 96 complex sounds of 500ms grouped in four categories of human voice, macaque vocalizations, marmoset vocalizations, and complex non-vocal sounds (more details in Bodin et al., 2021).

Main voice stimuli. The main stimulus set consisted of brief human voice sounds sampled from the Common Voice dataset (Ardila et al., 2020). Stimuli were organized into four main category levels: language (English, French, Spanish, Deutch, Polish, Portuguese, Russian, Chinese), gender (female/male), age (young/adult; young: teenagers and twenties; adult: thirties to sixties included) and identity (S1: 135 identities; S2: 142 identities; S3: 128 identities; ~44 samples per identity). Throughout the manuscript, the term 'gender' rather than 'sex' was utilized in reference to the demographic information obtained from the participants of the Common Voice dataset (Ardila et al., 2020), as it was the terminology employed in the survey ('male/female/other'). Stimulus sets were different for each participant and the number of stimuli per set also varied slightly (number of unique stimuli: Participant 1, N=6150; Participant 2, N=6148; Participant 3, N=5123). For each participant, six stimuli were selected randomly among the sounds having a higher energy (as measured by the amplitude envelope reaching an arbitrary threshold, likely corresponding to vowels) and were repeated extensively (60 times), to improve the performance of the brain decoding (VanRullen & Reddy, 2019; Horikawa & Kamitani, 2017; Chang et al., 2019); these will be called the "repeated" stimuli

hereafter, the remaining stimuli were presented twice. The third participant attended 5 BrainVoice sessions instead of 12, one BrainVoice session corresponding to 1030 stimuli (1024 unique stimuli and 6 'test' stimuli). Specifically, 5270 stimuli were presented to the third participant instead of ~12,000 for the two others. Among these 5270 stimuli, 5120 unique stimuli were presented once, as for the two other participants, 6 'test' stimuli were presented 25 times (150 trials). The dataset was fully balanced, with an equal number of samples for each combination of language, gender, age, and identity. Furthermore, to minimize potential adaptation effects, the stimuli were also balanced within each run according to these categories, and identity was balanced across sessions.

All stimuli of the main set were resampled at 24414 Hz and adjusted in duration (250 ms). For each stimulus, a fade-in and a fade-out were applied with a 15 ms cosine ramp to their onset and offset, and were normalized by dividing the root mean square amplitude. During fMRI sessions, stimulus presentations were controlled using custom Matlab scripts (Mathworks, Natick, MA, USA) interfaced with an RM1 Mobile Processor (Tucker-David Technologies, Alachua, USA). The auditory stimuli were delivered pseudo-randomly through MRI-compatible earphones (S14, SensiMetrics, USA) with no filtering and at a comfortable sound pressure level of around 85 dB SPL that allowed for clear and intelligible listening.

Computational models

We used two computational models to learn representational space for voice signals: Linear Autoencoder (LIN) and Deep Variational Autoencoder (VAE; Kingma & Welling., 2014). Both are encoder-decoder models that are learnt to reproduce at their output their input while going through a low-dimensional representation space usually called latent space (that we will call *voice latent space* since they are learnt on voice data). The autoencoders were trained on a dataset of 182K sounds from the Common Voice dataset (Ardila et al., 2020), balanced in gender, language and identity to reduce the bias in the synthesis (Gutierrez et al., 2021). Both models operate on sounds which were represented as spectrograms that we describe below. These representations were tested in all the encoding/decoding and RSA analyses.

Spectrograms

We used amplitude spectrograms as input for the models that we describe below. Short-term Fourier transforms of the waveform were computed using a sliding window of length 50 ms with a hop size of 12.5 ms (hence an overlap of 37.5 ms) and applying a Hamming window of size 800 samples before computing the Fourier transform of each slice. Only the magnitude of the spectrogram was kept and the phase of the complex representation was removed. At the end, a 250 ms sound is represented by a 21×401 matrix with 21 time steps and 401 frequency bins.

We used a custom code based on $numpy.fft$ package (Harris et al., 2020). The size and the overlap between the sliding windows of the spectrogram were chosen to conform with the uncertainty principle between time and frequency resolution. The main constraint was to find a trade-off between accurate phase reconstruction with the Griffin & Lim algorithm (1983) and a reasonable size of the spectrogram.

Deep neural network

We designed a deep variational autoencoder (VAE; Kingma & Welling, 2014) of 15 layers with an intermediate hidden representation of 128 neurons that we refer to as the *voice latent space* (VLS). In an autoencoder model, the two sub-network components, the *Encoder* and the *Decoder*, are jointly learned on complementary tasks (Figure 1a). The Encoder network (noted Enc hereafter; 7 layers) learns to map an input, s (a spectrogram of a sound), onto a (128-dimensional) *voice latent space* representation (z; in blue in the middle of Figure 1a), while the Decoder (noted Dec hereafter; 7 layers) aims at reconstructing the spectrogram s from z. The learning objective of the full model is to make the output spectrogram $Dec(Enc(s))$ as close as possible to the original one s. This reconstruction objective is defined as the L2 loss, $||Dec(Enc(s)) - s||^2$. The parameters of the Encoder and of the Decoder are jointly learned

using gradient descent to optimize the average L2 loss computed on the training set $\sum_{s \in Training\ Set} ||Dec(Enc(s)) - s||^2$. We trained this DNN on the Common Voice dataset (Ardila et al., 2020) according to VAE learning procedure (as explained in Kingma & Welling., 2019) until convergence (network architecture and particularities of the training procedure are provided in Supplementary Table 1), using the PyTorch python package (Paszke et al., 2019). Before feeding the spectrograms to the autoencoder, we standardized each of the 401 frequency bands separately by centering all the data corresponding to each frequency band at every time step in all spectrograms, which involved removing their mean, and dividing by their standard deviation. This separate standardization of frequency bands resulted in a smaller reconstruction error compared to standardizing across all the bands.

Linear autoencoder

We trained a linear autoencoder on the same dataset (described above) to serve as a linear baseline. Both the *Encoder* and the *Decoder* networks consisted of a single fully-connected layer, without any activation functions. Similar to the VAE, the latent space obtained from the *Encoder* was a 128-dimensional vector. The parameters of both the *Encoder* and of the *Decoder* were jointly learned using gradient descent to optimize the average L2 loss computed on the training set.

Neuroimaging data acquisition

Participants were scanned using a 3 Tesla Prisma scanner (Siemens Healthcare, Erlangen, Germany) equipped with a 64-channel receiver head-coil. Their movements were monitored during the acquisition using the software FIRMM (Dosenbach et al., 2017). The whole-head high-resolution structural scan acquired during the first session was a T1-weighted multi-echo MPRAGE (MEMPRAGE) (TR = 2.5 s, TE = 2.53, 4.28, 6.07, 7.86 ms, TI=1000 ms flip angle: 8°, matrix size = 208 × 300 × 320; resolution 0.8 × 0.8 × 0.8 mm^3, acquisition time: 8min22s). Lower resolution scans acquired during all other sessions were T1-weighted MPRAGE scans (TR = 2.3 s, TE = 2.88 ms, TI=900ms, flip angle: 9°, matrix size = 192 × 240 × 256; resolution 1 × 1 × 1 mm^3, sparse sampling with 2.8 times undersampling and compressed sensing reconstruction, acquisition time: 2min37). Functional imaging was performed using an EPI sequence (multiband factor = 5 , TR = 462 ms, TE = 31.2 ms, flip angle: 45°, matrix size = 84 × 84 × 35, resolution 2.5 × 2.5 × 2.5 mm^3). Functional slices were oriented parallel to the lateral sulci with a z-axis coverage of 87.5 mm, allowing it to fully cover both the TVAs (Pernet et al., 2015) and the FVAs (Aglieri et al., 2018). The physiological signals (heart rate and respiration) were measured with Siemens' external sensors.

Pre-processing of neuroimaging data and general linear modeling

Tissue segmentation and brain extraction were performed on the structural scans using the default segmentation procedure of SPM 12 (Ashburner et al., 2012). The preprocessing of the BOLD responses involved correcting motion, registering inter-runs, detrending and smoothing the data. Each functional volume was realigned to a reference volume taken from a steady period in the session that was spatially the closest to the average of all sessions. Transformation matrices between anatomical and functional data were computed using boundary-based registration (FSL; Smith et al., 2004). The data were respectively detrended and smoothed using the *nilearn* functions $clean_img$ and $smooth_img$ (kernel size of 3mm) (Abraham et al., 2014), resulting in the matrix $Y \in R^{S \times V}$, with S the number of scans and V the number of voxels.

A first general linear model (GLM) was fit to regress out the noise by predicting Y from a "denoised" design matrix composed of $R = 38$ regressors of nuisance (Figure 2-S2). These regressors of nuisance, also called covariates of no interest, included: 6 head motion parameters (3 variable for the translations, 3 variables for the rotations); 18 'RETROICOR' regressors (Glover et al., 2000) using the *TAPAS PhysIO* package (Kasper et al., 2017) (with the hyperparameters set as specified in Snoek et al.) were computed from the physiological signals; 13 regressors modeling slow artifactual trends (sines and cosines, cut frequency of

the high-pass filter = 0.01 Hz); and a confound-mean predictor. The design matrix was convolved with a hemodynamic response function (HRF) with a peak at 6 s and an undershoot at 16 s (Glover et al., 1999), we note the convolved design matrix as $X_d \in R^{S \times R}$. The "denoised" GLM's parameters $\beta_d \in R^{R \times V}$ were optimized to minimize the amplitude of the residual $\beta_d = argmin_{\beta \in R^{R \times V}} \; || \; Y \; - X_d \, \beta \; ||^2$. We used a lag-1 autoregressive model (ar(1)) to model the temporal structure of the noise (Friston et al., 2002). The *denoised* BOLD signal Y_d was then obtained from the original one according to $Y_d = Y - (X_d \, \beta_d) \in R^{S \times V}$.

A second "stimulus" GLM model was used to predict the denoised responses for each stimulus based on the denoised betas form the first GLM using a design matrix $X_s \in R^{S \times (N_S + 1)}$ (which was convolved with a hemodynamic response function, HRF as above) and a parameters matrix $\beta_s \in R^{(N_S+1) \times V}$ where N_S stands for the number of stimuli. The last row (resp. column) of β_s (resp. X_s) stands for a silence condition. Again, β_s was learned to minimize the residual $\beta_s = argmin_{\beta \in R^{(N_S+1) \times V}} \; || \, Y_d - X_s \, \beta \; ||^2$. Once learned, each of the first N_s line of β_s was corrected by subtracting the *(N_s+1)th* line, yielding the contrast maps for stimuli $\tilde{\beta}_s \in R^{N_S \times V}$. We note hereafter $\tilde{\beta}_s[i,:] \in R^V$ the contrast map for a given stimulus, it is the *i th* line of $\tilde{\beta}_s$.

A third "identity" GLM was fit to predict the BOLD responses of each voice speaker identity, using a design matrix $\beta_i \in R^{(N_i+1) \times V}$ and a design matrix $X_i \in R^{S \times (N_i + 1)}$ (which was again convolved with a hemodynamic response function, HRF) where N_s stands for the number of unique speakers. Again the last row/column in β_i and X_i stands for the silent condition. β_i is learned to minimize the residual $\beta_i = argmin_{\beta \in R^{(N_i+1) \times V}} ||Y_d - X_i \, \beta \; ||^2$ (Figure 2-S1a). Again, the final speaker contrast maps were obtained by contrasting (i.e., subtracting) the regression coefficients in a row of β_i with the silence condition (last row; Figure 2-S1a), yielding $\tilde{\beta}_i \in R^{N_s \times V}$. Here the *jth* row of $\tilde{\beta}_i$, $\tilde{\beta}_i[j,:] \in R^V$, represents the amplitude of the BOLD response of the contrast map for speaker *j* (i.e. to all the stimuli from this speaker).

A fourth "localizer" GLM model was used to predict the denoised BOLD responses of each sound category from the *Voice localizer stimuli* presented above. The procedure was similar as described for the two previous GLM models. Once the GLM was learned, we contrasted the human voice category with the other sound categories in order to localize for each participant the posterior Temporal Voice Area (pTVA), medial Temporal Voice Area (mTVA) and anterior Temporal Voice Area (aTVA) in each hemisphere. The center of each TVA corresponded to the local maximum of the voice > non-voice t-map whose coordinates were the closest to the TVAs reported in (Pernet et al., 2015). The analyses were carried on for each region of interest (ROI) of each hemisphere.

Additionally, we defined for each participant the primary auditory cortex (A1) as the maximum value of the probabilistic map (non-linearly registered to each participant functional space) of Heschl's gyri provided with the MNI152 template (Penhune et al., 1996), intersected with the sound vs silence contrast map.

Identity-based and stimulus-based representations

We performed analyses either at the stimulus level, e.g. predicting the neural activity of a participant listening to a given *stimulus* ($\tilde{\beta}_s$'s lines) from the *voice latent space* representation of this stimuli, or at the speaker identity level, e.g. predicting the average neural activity in response to stimuli of a given speaker *identity* ($\tilde{\beta}_i$'s lines) from this speaker's *voice latent space* representation. The identity-based analyses were used for the characterization of the *voice latent space* (Figure 1), the brain encoding (Figure 2), and the representational similarity analysis (Figure 3), while the stimulus-based analyses were used for the brain decoding analyses (Figure 4, 5).

We conducted stimulus-based analyses to examine the relationship between stimulus contrast maps in neural activity ($\tilde{\beta}_s$) and the encodings of individual stimulus spectrograms computed

by the encoder of an autoencoder model (either linear or deep variational autoencoder) on the computational side. We will note ${z_s}^{lin} \in R^{N_s \times 128}$ the encodings of stimuli by the LIN model and ${z_s}^{vae} \in R^{N_s \times 128}$ the encodings of stimuli computed by the VAE model. The encoding of the kth stimuli by one of these models is the kth row of the corresponding matrix and it is noted as ${z_s}^{model}[k,:]$.

For identity-based analyses we studied relationships between identity contrast maps in $\tilde{\beta}_i$ on the neural activity side, and an encoding of speaker identity in the VLS implemented by an autoencoder model (LIN or VAE) on the computational side, e.g. we note ${z_i}^{vae}[j]$ the representation of speaker *j* as computed by the *vae* model. We chose to define a speaker identity-based representation as the average of a set of sample-based representations for stimuli from this speaker, e.g. ${z_i}^{model}[j] = 1/|S_j| \sum_{k \in S_j} {z_s}^{model}[k,:]$ where S_j stands for the set of stimuli by speaker *j* and *model* stands for *vae* or *lin*. Averaging in the *voice latent space* is expected to be much more powerful and relevant than averaging in the input space spectrograms (VanRullen & Reddy, 2019).

Characterization of the autoencoder latent space

We characterized the organization of the *voice latent space* (VLS) and of the features computed by the linear autoencoder (LIN) by measuring through classification experiments the presence of information about speaker's gender, age, and identity in the representations learned by these models.

We first computed the speaker's identity *voice latent space* representations for each of the 405 speakers in the main voice dataset (135+142+128 see *Stimuli* section) as explained above.

Next we used these speakers' *voice latent space* representation to investigate if the gender, age, identity were encoded in the VLS. To do so, we divided the data into separate train and test sets and learned classifiers to predict gender, age, or identity from the train set. The balanced (to avoid the small effects associated with unbalanced folds) accuracy of the classifiers was then evaluated on the test set. The higher the performance on the test set, the more confident we are that the information is encoded in the VLS. More specifically for each task (gender, age, identity), we trained a Logistic Regression classifier (linear regularized logistic regression; L2 penalty, tol=0.0001, fit_intercept=True, intercept_scaling=1, max_iter=100) using the scikit-learn python package (Pedregosa et al., 2018).

We performed 5 train-test splits with 80% of the data in the training and 20% in the test set. For each split we used 5-fold cross-validation on the training set to select the optimal value for the regularization hyperparameter C (searching between 10 values logarithmically spaced on the interval [-3, +3]). We then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results were then averaged over 5 folds. Note that data were systematically normalized with a scaler fitted on the training set. We used a robust scaling strategy for these experiments (removing the median, then scaling to the quantile range; 25th quantile and 75th quantile) which occurs to be more relevant with a small training set.

To investigate how speaker identity information is encoded in the latent space representations of speakers' voices, we computed speaker identity *voice latent space* representations by averaging 20 stimulus-based representations, in order to obtain a limited amount of data per identity that could be distributed across training and test datasets.

We tested whether the mean of the distribution of accuracy scores obtained for 5 folds was significantly above chance level using Wilcoxon signed-rank test.

Brain encoding
We performed encoding experiments on identity-based representations for each of the three participants (Figure 2). For each participant we learn a regularized linear regression that predicts a speaker-based neural activity, e.g. the jth speaker's contrast map $\tilde{\beta}_i[j] \in R^V$, from this speaker's voice latent space representation, that we note $z_i{}^{model}[j] \in R^{128}$ (Figure 2a), where i is the voxel index. We carried out these regression analyses for each ROI (A1, pTVA, mTVA, aTVA) in each hemisphere and participant, independently.

The regression model parameters $\widehat{W}_{encod} \in R^{128\times V}$ were learned according to:

$$\widehat{W}_{encod} = argmin_{W_{encod}\in R^{128\times V}} \sum_{j=1..N_i} (z_i{}^{model}[j] \times W_{encod} \quad - \tilde{\beta}_i[j])^2 + \lambda \|W_{encod}\|^2$$

where λ is a hyperparameter tuning the optimal tradeoff between the data fit and the penalization terms above. We used the ridge regression with built-in cross-validation as implemented as $RidgeCV$ in the scikit-learn library (Pedregosa et al., 2018).

The statistical significance of each result was assessed using the following procedure: We repeated the following experiment 20 times with different random seeds. Each time, we performed 5 train-test splits with 80% of the data in the training and 20% in the test set. For each split we used RidgeCV (relying on leave-one-out) on the training set to select the optimal value for the hyperparameter λ (searching between 10 values logarithmically spaced on the interval $[10^{-1}; 10^{8}]$). Following standard practice in machine learning, we then computed the generalization performance on the test set of the model trained on the full training set with the best hyperparameter value. Reported results are then averaged over 20 experiments. Note that here again with small training sets data were systematically normalized in each experiment using robust scaling.

The evaluation relied on the 'brain score'-inspired procedure (Schrimpf et al., 2018) which evaluates the performance of the ridge regression with a Pearson's correlation score. Correlations between measured neural activities $\tilde{\beta}_i$ and predicted ones $z_i{}^{\widehat{model}} * W_{encod}$ were computed for each voxel and averaged over repeated experiments (folds and seeds) yielding one correlation value for every voxel and for every setting. The significance of the results was assessed with one-sample t-tests for the Fisher z-transformed correlation scores (3 x participants x 2 hemispheres x V voxels). For each region of interest, the scores are reported across participants and hemispheres (Figure 2b). The exact same procedure was followed for the LIN modeling, and for the Wav2Vec and HuBERT models (cf. Fig2S3).

In order to determine which of the two feature spaces (VLS, LIN) and which of the two ROI (A1, TVAs) yielded the best prediction of neural activity, we compared the means of distributions of correlations coefficients using a mixed ANOVA performed on the Fisher z-transformed coefficients (dependent variable: correlation; between factor: ROI; repeated measurements: Feature; between-participant identifier: voxel).

For each ROI, we then used t-tests to perform post-hoc contrasts for the VLS-LIN difference in brain encoding performance (comparison tests in Figure 2b; Supplementary Table 4). We finally conducted two-sample t-tests between the brain encoding model's scores trained to predict A1 and those trained to predict temporal voice areas to test the significance of the A1-TVAs difference (Supplementary Table 5).

The statistical tests were all performed using the *pingouin* python package (Vallat, 2018).

Representational similarity analysis
The RSA analyses were carried out using the package *rsatoolbox* (Schütt et al., 2021; https://github.com/rsagroup/rsatoolbox). For each participant, region of interest and

hemisphere, we computed the cerebral Representational Dissimilarity Matrix (RDM) using Pearson's correlation between the speaker identity-specific response patterns of the GLM estimates $\tilde{\beta}_i$ (Walther et al., 2016) (Figure 3a). The model RDMs were built using cosine distance (Xing et al., 2015; Bhattacharya et al., 2017; Wang et al., 2018), capturing speaker pairwise feature differences predicted by the computational models LIN and the VLS (Figure 3a). For greater comparability with the rest of the analyses described here, the GLM estimates and the computational models' features were first normalized using robust scaling. We computed the Spearman correlations coefficients between the brain RDMs for each ROI, and the two model's RDMs (Figure 3b). We assessed the significance of these brain-model correlation coefficients within a permutation-based 'maximum statistics' framework for multiple comparison correction (one-tailed inference; N permutations = 10,000 for each test; permutation of rows and columns of distance matrices, see Giordano et al., 2023 and Maris & Oostenveld, 2007; see Figure 3b). We evaluated the VLS-LIN difference using a two-way repeated-measures ANOVA on the Fisher z-transformed Spearman correlation coefficients (dependent variable: correlation; within factors: ROI and Feature; participant identifier: participant hemisphere pair). The same permutation framework was also used to assess the significance of the difference between the RSA correlation for the VLS and LIN models.

Brain decoding

Brain decoding was investigated at the stimulus level. The stimuli's voice latent space representations $z_s^{model} \in R^{N \times 128}$ and voice samples' contrast maps $\tilde{\beta}_s \in R^{N \times V}$ were divided into train and test splits, normalized across voice samples using robust scaling, and then fit into the training set. For every participant and each ROI, we trained a L_2-regularized linear model $W \in R^{V \times 128}$ to predict the voice samples' latent vectors from the voice samples' contrast maps (Figure 4a). The hyperparameter selection and optimization were made similarly as in the Brain encoding scheme. Training was performed on non-repeated stimuli (see Stimuli section). We then used the trained models to predict for each participant the 6 repeated stimuli that were the most presented. Waveforms were estimated starting from the reconstructed spectrograms using the Griffin-Lim phase reconstruction algorithm (Griffin & Lim, 1983).

We then used classifier analyses to assess the presence of voice information (gender, age, speaker identity) in the reconstructed latent representations (i.e., the latent representation predicted from the brain activity of a participant listening to a specific stimulus) (Figure 5a, b, c). To this purpose, we first trained linear classifiers to categorize the training voice stimuli (participant 1, N = 6144; participant 2, N = 6142; participant 3, N = 5117; total, N = 17403) by gender (2 classes), age (2 classes) or identity (17 classes) based on VLS coordinates. Secondly, we used the previously trained classifiers to predict the identity information based on the VLS derived from the brain responses of the 18 Test voice stimuli (3 participants x 6 stimuli). We first tested using one-sample t-tests that the mean of the distribution of accuracy scores obtained across random classifier initializations of classifiers (2 hemispheres x 20 seeds = 40) was significantly above chance level, for each category, ROI and model. We then evaluated the difference in performance at preserving identity-related information depending on the model or ROI via two-way ANOVAs (dependent variable: accuracy; between factors: Feature and ROI). We performed post-hoc planned paired t-tests between each model pair to test the significance of the VLS-LIN difference. Two-sample t-tests were finally used to test the significance of the A1-TVAs difference.

Listening tests

We recruited 13 participants through the online platform Prolific (www.prolific.co) for a series of online behavioral experiments. All participants reported having normal hearing. The purpose of these experiments was to evaluate how well voice identity information and naturalness are preserved in fMRI-based reconstructed voice excerpts. In the main session, participants carried out 4 tasks, in the following order: 'speaker discrimination' (~120 min), 'perceived naturalness' (~30 min), 'gender categorization' (~30 min), 'age categorization' (~30

min). The experiment lasted 3 hours and 35 minutes, and each participant was paid £48. 12 participants performed the speaker discrimination task, and all participants performed the other tasks.

Prior to the main experiment session, participants carried out a short loudness-change detection task to ensure that they wore headphones and that they were attentive and properly set up for the main experiment (Woods et al., 2017). On each of the 12 trials, participants heard 3 tones and were asked to identify which tone was the least loud by clicking on one of 3 response buttons: 'First', 'Second', or 'Third'. Participants were admitted to the main experiment only if they achieved perfect performance in this task. We additionally refined the participant pool by excluding those who performed badly on the original stimuli, by retaining only the subjects whose performance was above the 25th percentile of accuracy. (gender and age categorizations: as all participants performed well (Figure 5d, e, red dotted lines); speaker discrimination: 9/12 participants performed above the threshold of 64%).

The next three tasks were each carried out on the same set of 342 experimental stimuli, each presented on a different trial: 18 original stimuli, 36 stimuli reconstructed directly from the LIN and the VLS models, and 18 stimuli x 2 models x 4 regions of interest x 2 hemispheres= 288 brain-reconstructed stimuli.

In the 'perceived naturalness' task, participants were asked to rate how natural the voice sounded on a scale ranging from 'Not at all natural' to 'Highly natural' (i.e., similar to a real recording), and were instructed to use the full range of the scale.

During the 'gender categorization' task, participants categorized the gender by clicking on a 'Female' or 'Male' button.

Finally, in the 'age categorization' task, participants categorized the age of the speaker by clicking on a 'Younger' or 'Older' button.

In the 'speaker discrimination' task, participants carried out 684 trials (342 experimental stimuli x 2) with short breaks in between. On each trial, they were presented with 2 short sound stimuli, one after the other, and participants had to indicate whether they were from the same speaker or not. The speech material was selected randomly and was different between two stimuli.

To evaluate the performance of the participants, we first conducted one-sample t-tests to examine whether the mean accuracy score calculated from their responses was significantly higher than the chance level for each model and ROI. Next, we used two-way repeated-measures ANOVAs to assess the variation in participants' performances in identifying identity-related information (dependent variable: accuracy; between-participant factors: Feature and ROI). To determine the statistical significance of the VLS-LIN difference, we carried out post-hoc planned paired t-tests between each model pair. Finally, we employed two-sample t-tests to evaluate the statistical significance of the A1-TVAs difference.

Data and code availability

The preprocessed data and codes are publicly available on Zenodo: https://doi.org/10.5281/zenodo.15797934.

References

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. 2014. "Machine Learning for Neuroimaging with Scikit-Learn." *Frontiers in Neuroinformatics* 8.

Aglieri, Virginia, Bastien Cagna, Lionel Velly, Sylvain Takerkart, and Pascal Belin. 2021. "FMRI-Based Identity Classification Accuracy in Left Temporal and Frontal Regions Predicts Speaker Recognition Performance." *Scientific Reports* 11(1):489. doi: 10.1038/s41598-020-79922-7.

Akbari, Hassan, Bahar Khalighinejad, Jose L. Herrero, Ashesh D. Mehta, and Nima Mesgarani. 2019. "Towards Reconstructing Intelligible Speech from the Human Auditory Cortex." *Scientific Reports* 9(1):874. doi: 10.1038/s41598-018-37359-z.

Ardila, Rosana, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. "Common Voice: A Massively-Multilingual Speech Corpus."

Ashburner, John. 2012. "SPM: A History." *NeuroImage* 62(2):791–800. doi: 10.1016/j.neuroimage.2011.10.025.

Barbero, Francesca M., Roberta P. Calce, Siddharth Talwar, Bruno Rossion, and Olivier Collignon. 2021. "Fast Periodic Auditory Stimulation Reveals a Robust Categorical Response to Voices in the Human Brain." *eNeuro* 8(3):ENEURO.0471-20.2021. doi: 10.1523/ENEURO.0471-20.2021.

Belin, Pascal, Patricia E. G. Bestelmeyer, Marianne Latinus, and Rebecca Watson. 2011. "Understanding Voice Perception: Understanding Voice Perception." *British Journal of Psychology* 102(4):711–25. doi: 10.1111/j.2044-8295.2011.02041.x.

Belin, Pascal, Clémentine Bodin, and Virginia Aglieri. 2018. "A 'Voice Patch' System in the Primate Brain for Processing Vocal Information?" *Hearing Research* 366:65–74. doi: 10.1016/j.heares.2018.04.010.

Belin, Pascal, Shirley Fecteau, and Catherine Bédard. 2004. "Thinking the Voice: Neural Correlates of Voice Perception." *Trends in Cognitive Sciences* 8(3):129–35. doi: 10.1016/j.tics.2004.01.008.

Belin, Pascal, and Robert J. Zatorre. 2003. "Adaptation to Speaker's Voice in Right Anterior Temporal Lobe:" *NeuroReport* 14(16):2105–9. doi: 10.1097/00001756-200311140-00019.

Belin, Pascal, Robert J. Zatorre, Philippe Lafaille, Pierre Ahad, and Bruce Pike. 2000. "Voice-Selective Areas in Human Auditory Cortex." *Nature* 403(6767):309–12. doi: 10.1038/35002078.

Bhattacharya, Gautam, Jahangir Alam, and Patrick Kenny. 2017. "Deep Speaker Embeddings for Short-Duration Speaker Verification." Pp. 1517–21 in *Interspeech 2017*. ISCA.

Blank, Helen, Nuri Wieland, and Katharina von Kriegstein. 2014. "Person Recognition and the Brain: Merging Evidence from Patients and Healthy Individuals." *Neuroscience & Biobehavioral Reviews* 47:717–34. doi: 10.1016/j.neubiorev.2014.10.022.

Bodin, Clémentine, Régis Trapeau, Bruno Nazarian, Julien Sein, Xavier Degiovanni, Joël Baurberg, Emilie Rapha, Luc Renaud, Bruno L. Giordano, and Pascal Belin. 2021.

"Functionally Homologous Representation of Vocalizations in the Auditory Cortex of Humans and Macaques." *Current Biology* S0960982221011477. doi: 10.1016/j.cub.2021.08.043.

Capilla, A., P. Belin, and J. Gross. 2013. "The Early Spatio-Temporal Correlates and Task Independence of Cerebral Voice Processing Studied with MEG." *Cerebral Cortex* 23(6):1388–95. doi: 10.1093/cercor/bhs119.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Rémi King. 2022. "Deep Language Algorithms Predict Semantic Comprehension from Brain Activity." *Scientific Reports* 12(1):16327. doi: 10.1038/s41598-022-20460-9.

Caucheteux, Charlotte, Alexandre Gramfort, and Jean-Rémi King. 2023. "Evidence of a Predictive Coding Hierarchy in the Human Brain Listening to Speech." *Nature Human Behaviour* 1–12. doi: 10.1038/s41562-022-01516-2.

Caucheteux, Charlotte, and Jean-Rémi King. 2022. "Brains and Algorithms Partially Converge in Natural Language Processing." *Communications Biology* 5(1):1–10. doi: 10.1038/s42003-022-03036-1.

Chang, Nadine, John A. Pyles, Austin Marcus, Abhinav Gupta, Michael J. Tarr, and Elissa M. Aminoff. 2019. "BOLD5000, a Public fMRI Dataset While Viewing 5000 Visual Images." *Scientific Data* 6(1):49. doi: 10.1038/s41597-019-0052-3.

Charest, I., C. Pernet, M. Latinus, F. Crabbe, and P. Belin. 2013. "Cerebral Processing of Voice Gender Studied Using a Continuous Carryover fMRI Design." *Cerebral Cortex* 23(4):958–66. doi: 10.1093/cercor/bhs090.

Charest, Ian, Cyril R. Pernet, Guillaume A. Rousselet, Ileana Quiñones, Marianne Latinus, Sarah Fillion-Bilodeau, Jean-Pierre Chartrand, and Pascal Belin. 2009. "Electrophysiological Evidence for an Early Processing of Human Voices." *BMC Neuroscience* 10(1):127. doi: 10.1186/1471-2202-10-127.

Chen, Zijiao, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, and Juan Helen Zhou. 2023. "Seeing Beyond the Brain: Conditional Diffusion Model with Sparse Masked Modeling for Vision Decoding." Pp. 22710–20 in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE.

Cowen, Alan S., Marvin M. Chun, and Brice A. Kuhl. 2014. "Neural Portraits of Perception: Reconstructing Face Images from Evoked Brain Activity." *NeuroImage* 94:12–22. doi: 10.1016/j.neuroimage.2014.03.018.

Dado, Thirza, Yağmur Güçlütürk, Luca Ambrogioni, Gabriëlle Ras, Sander Bosch, Marcel van Gerven, and Umut Güçlü. 2022. "Hyperrealistic Neural Decoding for Reconstructing Faces from fMRI Activations via the GAN Latent Space." *Scientific Reports* 12(1):141. doi: 10.1038/s41598-021-03938-w.

Défossez, Alexandre, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. "Decoding Speech Perception from Non-Invasive Brain Recordings." *Nature Machine Intelligence* 5(10):1097–1107. doi: 10.1038/s42256-023-00714-5.

Diedrichsen, Jörn, and Nikolaus Kriegeskorte. 2017. "Representational Models: A Common Framework for Understanding Encoding, Pattern-Component, and Representational-Similarity Analysis." *PLOS Computational Biology* 13(4):e1005508. doi: 10.1371/journal.pcbi.1005508.

Dosenbach, Nico U. F., Jonathan M. Koller, Eric A. Earl, Oscar Miranda-Dominguez, Rachel L. Klein, Andrew N. Van, Abraham Z. Snyder, Bonnie J. Nagel, Joel T. Nigg, Annie L. Nguyen, Victoria Wesevich, Deanna J. Greene, and Damien A. Fair. 2017. "Real-Time Motion Analytics during Brain MRI Improve Data Quality and Reduce Costs." *NeuroImage* 161:80–93. doi: 10.1016/j.neuroimage.2017.08.025.

Friston, K. J., D. E. Glaser, R. N. A. Henson, S. Kiebel, C. Phillips, and J. Ashburner. 2002. "Classical and Bayesian Inference in Neuroimaging: Applications." *NeuroImage* 16(2):484–512. doi: 10.1006/nimg.2002.1091.

Friston, K. J., A. P. Holmes, K. J. Worsley, J. P. Poline, C. D. Frith, and R. S. J. Frackowiak. 1994. "Statistical Parametric Maps in Functional Imaging: A General Linear Approach." *Human Brain Mapping* 2(4):189–210. doi: 10.1002/hbm.460020402.

Gaziv, Guy, Roman Beliy, Niv Granot, Assaf Hoogi, Francesca Strappini, Tal Golan, and Michal Irani. 2022. "Self-Supervised Natural Image Reconstruction and Large-Scale Semantic Classification from Brain Activity." *NeuroImage* 254:119121. doi: 10.1016/j.neuroimage.2022.119121.

Giordano, Bruno L., Michele Esposito, Giancarlo Valente, and Elia Formisano. 2023. "Intermediate Acoustic-to-Semantic Representations Link Behavioral and Neural Responses to Natural Sounds." *Nature Neuroscience* 1–9. doi: 10.1038/s41593-023-01285-9.

Glover, G. H., T. Q. Li, and D. Ress. 2000. "Image-Based Method for Retrospective Correction of Physiological Motion Effects in fMRI: RETROICOR." *Magnetic Resonance in Medicine* 44(1):162–67. doi: 10.1002/1522-2594(200007)44:1<162::aid-mrm23>3.0.co;2-e.

Glover, Gary H. 1999. "Deconvolution of Impulse Response in Event-Related BOLD fMRI1." *NeuroImage* 9(4):416–29. doi: 10.1006/nimg.1998.0419.

Griffin, D. and Jae Lim. 1983. "Signal Estimation from Modified Short-Time Fourier Transform." Pp. 804–7 in *ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 8. Boston, MASS, USA: Institute of Electrical and Electronics Engineers.

Güçlü, Umut, Jordy Thielen, Michael Hanke, M. A. J. van Gerven, and Marcel A. J. van Gerven. 2016. "Brains on Beats." *In Proceedings of the International Conference on Neural Information Processing Systems* 2101–9. doi: null.

Gutierrez, Miren. 2021. "Algorithmic Gender Bias and Audiovisual Data: A Research Agenda." *International Journal of Communication* 15:439–61.

Han, Kuan, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. 2019. "Variational Autoencoder: An Unsupervised Model for Encoding and Decoding fMRI Activity in Visual Cortex." *NeuroImage* 198:125–36. doi: 10.1016/j.neuroimage.2019.05.039.

Harris, Charles R., K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. "Array Programming with NumPy." *Nature* 585(7825):357–62. doi: 10.1038/s41586-020-2649-2.

Higgins, Irina, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. 2021. "Unsupervised Deep Learning Identifies Semantic Disentanglement in Single Inferotemporal Face Patch Neurons." *Nature Communications* 12(1):6456. doi: 10.1038/s41467-021-26751-5.

Horikawa, Tomoyasu, and Yukiyasu Kamitani. 2017. "Generic Decoding of Seen and Imagined Objects Using Hierarchical Visual Features." *Nature Communications* 8(1):15037. doi: 10.1038/ncomms15037.

Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." *arXiv:1704.04861 [Cs]*.

Kasper, Lars, Steffen Bollmann, Andreea O. Diaconescu, Chloe Hutton, Jakob Heinzle, Sandra Iglesias, Tobias U. Hauser, Miriam Sebold, Zina-Mary Manjaly, Klaas P. Pruessmann, and Klaas E. Stephan. 2017. "The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data." *Journal of Neuroscience Methods* 276:56–72. doi: 10.1016/j.jneumeth.2016.10.019.

Kell, Alexander J. E., Daniel L. K. Yamins, Erica N. Shook, Sam V. Norman-Haignere, and Josh H. McDermott. 2018. "A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy." *Neuron* 98(3):630-644.e16. doi: 10.1016/j.neuron.2018.03.044.

Khaligh-Razavi, Seyed-Mahdi, and Nikolaus Kriegeskorte. 2014. "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation" edited by J. Diedrichsen. *PLoS Computational Biology* 10(11):e1003915. doi: 10.1371/journal.pcbi.1003915.

Kingma, Diederik P., and Max Welling. 2014. "Auto-Encoding Variational Bayes." *arXiv:1312.6114 [Cs, Stat]*.

Kingma, Diederik P., and Max Welling. 2019. "An Introduction to Variational Autoencoders." *Foundations and Trends® in Machine Learning* 12(4):307–92. doi: 10.1561/2200000056.

Kriegeskorte, Nikolaus. 2008. "Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience." *Frontiers in Systems Neuroscience*. doi: 10.3389/neuro.06.004.2008.

Kriegstein, Katharina V., and Anne-Lise Giraud. 2004. "Distinct Functional Substrates along the Right Superior Temporal Sulcus for the Processing of Voices." *NeuroImage* 22(2):948–55. doi: 10.1016/j.neuroimage.2004.02.020.

von Kriegstein, Katharina, Evelyn Eger, Andreas Kleinschmidt, and Anne Lise Giraud. 2003. "Modulation of Neural Responses to Speech by Directing Attention to Voices or Verbal Content." *Cognitive Brain Research* 17(1):48–55. doi: 10.1016/S0926-6410(03)00079-X.

Lavan, Nadine. 2023. "The Time Course of Person Perception From Voices: A Behavioral Study." *Psychological Science* 34(7):771–83. doi: 10.1177/09567976231161565.

Le, Lynn, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. 2022. "Brain2Pix: Fully Convolutional Naturalistic Video Frame Reconstruction from Brain Activity." *Frontiers in Neuroscience* 16.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521(7553):436–44. doi: 10.1038/nature14539.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., & Chang, E. F. (2023). Dissecting neural computations in the human auditory pathway using deep

neural networks for speech. *Nature Neuroscience*, *26*(12), 2213–2225. https://doi.org/10.1038/s41593-023-01468-4

Liu, Dong, Yue Li, Jianping Lin, Houqiang Li, and Feng Wu. 2020. "Deep Learning-Based Video Coding: A Review and a Case Study." *ACM Computing Surveys* 53(1):11:1-11:35. doi: 10.1145/3368405.

Maguinness, Corrina, Claudia Roswandowitz, and Katharina von Kriegstein. 2018. "Understanding the Mechanisms of Familiar Voice-Identity Recognition in the Human Brain." *Neuropsychologia* 116:179–93. doi: 10.1016/j.neuropsychologia.2018.03.039.

Maris, Eric, and Robert Oostenveld. 2007. "Nonparametric Statistical Testing of EEG- and MEG-Data." *Journal of Neuroscience Methods* 164(1):177–90. doi: 10.1016/j.jneumeth.2007.03.024.

Martin, F. N., and C. A. Champlin. 2000. "Reconsidering the Limits of Normal Hearing." *Journal of the American Academy of Audiology* 11(2):64–66.

Mcallister, Jan, Anne Potts, Kathryn Mason, and Geoffrey Marchant. 1994. "Word Duration in Monologue and Dialogue Speech." *Language and Speech* 37(4):393–405. doi: 10.1177/002383099403700404.

Millet, Juliette, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. 2022. *Toward a Realistic Model of Speech Processing in the Brain with Self-Supervised Learning*. arXiv:2206.01685. arXiv. doi: 10.48550/arXiv.2206.01685.

Mozafari, Milad, Leila Reddy, and Rufin VanRullen. 2020. "Reconstructing Natural Scenes from fMRI Patterns Using BigBiGAN." *2020 International Joint Conference on Neural Networks (IJCNN)* 1–8. doi: 10.1109/IJCNN48605.2020.9206960.

Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. 2017. "VoxCeleb: A Large-Scale Speaker Identification Dataset." Pp. 2616–20 in *Interspeech 2017*.

Naselaris, Thomas, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. 2011. "Encoding and Decoding in fMRI." *NeuroImage* 56(2):400–410. doi: 10.1016/j.neuroimage.2010.07.073.

Pasley, Brian N., Stephen V. David, Nima Mesgarani, Adeen Flinker, Shihab A. Shamma, Nathan E. Crone, Robert T. Knight, and Edward F. Chang. 2012. "Reconstructing Speech from Human Auditory Cortex." *PLOS Biology* 10(1):e1001251. doi: 10.1371/journal.pbio.1001251.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library."

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2018. "Scikit-Learn: Machine Learning in Python."

Penhune, V. B., R. J. Zatorre, J. D. MacDonald, and A. C. Evans. 1996. "Interhemispheric Anatomical Differences in Human Primary Auditory Cortex: Probabilistic Mapping and

Volume Measurement from Magnetic Resonance Scans." *Cerebral Cortex (New York, N.Y.: 1991)* 6(5):661–72. doi: 10.1093/cercor/6.5.661.

Pernet, Cyril R., Phil McAleer, Marianne Latinus, Krzysztof J. Gorgolewski, Ian Charest, Patricia E. G. Bestelmeyer, Rebecca H. Watson, David Fleming, Frances Crabbe, Mitchell Valdes-Sosa, and Pascal Belin. 2015. "The Human Voice Areas: Spatial Organization and Inter-Individual Variability in Temporal and Extra-Temporal Cortices." *NeuroImage* 119:164–74. doi: 10.1016/j.neuroimage.2015.06.050.

Petkov, Christopher I., Christoph Kayser, Thomas Steudel, Kevin Whittingstall, Mark Augath, and Nikos K. Logothetis. 2008. "A Voice Region in the Monkey Brain." *Nature Neuroscience* 11(3):367–74. doi: 10.1038/nn2043.

Roswandowitz, Claudia, Claudia Kappes, Hellmuth Obrig, and Katharina von Kriegstein. 2018. "Obligatory and Facultative Brain Regions for Voice-Identity Recognition." *Brain* 141(1):234–47. doi: 10.1093/brain/awx313.

Rupp, Kyle, Jasmine L. Hect, Madison Remick, Avniel Ghuman, Bharath Chandrasekaran, Lori L. Holt, and Taylor J. Abel. 2022. "Neural Responses in Human Superior Temporal Cortex Support Coding of Voice Representations." *PLOS Biology* 20(7):e3001675. doi: 10.1371/journal.pbio.3001675.

Santoro, Roberta, Michelle Moerel, Federico De Martino, Giancarlo Valente, Kamil Ugurbil, Essa Yacoub, and Elia Formisano. 2017. "Reconstructing the Spectrotemporal Modulations of Real-Life Sounds from fMRI Response Patterns." *Proceedings of the National Academy of Sciences* 114(18):4799–4804. doi: 10.1073/pnas.1617622114.

Schrimpf, Martin, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. "The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing." *Proceedings of the National Academy of Sciences* 118(45):e2105646118. doi: 10.1073/pnas.2105646118.

Schrimpf, Martin, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. *Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like? preprint*. Neuroscience. doi: 10.1101/407007.

Schütt, Heiko H., Alexander D. Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. 2021. "Statistical Inference on Representational Geometries."

Schweinberger, S. R., A. Herholz, and W. Sommer. 1997. "Recognizing Famous Voices: Influence of Stimulus Duration and Different Types of Retrieval Cues." *Journal of Speech, Language, and Hearing Research: JSLHR* 40(2):453–63. doi: 10.1044/jslhr.4002.453.

Smith, Stephen M., Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E. J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. 2004. "Advances in Functional and Structural MR Image Analysis and Implementation as FSL." *NeuroImage* 23:S208–19. doi: 10.1016/j.neuroimage.2004.07.051.

Snoek, Lukas, Maite M. van der Miesen, Tinka Beemsterboer, Andries van der Leij, Annemarie Eigenhuis, and H. Steven Scholte. 2021. "The Amsterdam Open MRI Collection, a Set of Multimodal MRI Datasets for Individual Difference Analyses." *Scientific Data* 8(1):85. doi: 10.1038/s41597-021-00870-6.

Trapeau, Régis, Etienne Thoret, and Pascal Belin. 2022. “The Temporal Voice Areas Are Not ‘Just’ Speech Areas.” *Frontiers in Neuroscience* 16:1075288. doi: 10.3389/fnins.2022.1075288.

Tuckute, Greta, Jenelle Feather, Dana Boebinger, and Josh H. McDermott. 2023. “Many but Not All Deep Neural Network Audio Models Capture Brain Responses and Exhibit Correspondence between Model Stages and Brain Regions.” *PLOS Biology* 21(12):e3002366. doi: 10.1371/journal.pbio.3002366.

Vallat, Raphael. 2018. “Pingouin: Statistics in Python.” *Journal of Open Source Software* 3(31):1026. doi: 10.21105/joss.01026.

VanRullen, Rufin, and Leila Reddy. 2019. “Reconstructing Faces from fMRI Patterns Using Deep Generative Neural Networks.” *Communications Biology* 2(1):193. doi: 10.1038/s42003-019-0438-y.

Walther, Alexander, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. 2016. “Reliability of Dissimilarity Measures for Multi-Voxel Pattern Analysis.” *NeuroImage* 137:188–200. doi: 10.1016/j.neuroimage.2015.12.012.

Wang, Xiaosha, Yangwen Xu, Yuwei Wang, Yi Zeng, Jiacai Zhang, Zhenhua Ling, and Yanchao Bi. 2018. “Representational Similarity Analysis Reveals Task-Dependent Semantic Influence of the Visual Word Form Area.” *Scientific Reports* 8(1):3047. doi: 10.1038/s41598-018-21062-0.

Wetzel, Sebastian J. 2017. “Unsupervised Learning of Phase Transitions: From Principal Component Analysis to Variational Autoencoders.” *Physical Review E* 96(2):022140. doi: 10.1103/PhysRevE.96.022140.

Woods, Kevin J. P., Max H. Siegel, James Traer, and Josh H. McDermott. 2017. “Headphone Screening to Facilitate Web-Based Auditory Experiments.” *Attention, Perception, & Psychophysics* 79(7):2064–72. doi: 10.3758/s13414-017-1361-2.

Wu, Michael C. K., Stephen V. David, and Jack L. Gallant. 2006. “COMPLETE FUNCTIONAL CHARACTERIZATION OF SENSORY NEURONS BY SYSTEM IDENTIFICATION.” *Annual Review of Neuroscience* 29(1):477–505. doi: 10.1146/annurev.neuro.29.051605.113024.

Xing, Chao, Dong Wang, Chao Liu, and Yiye Lin. 2015. “Normalized Word Embedding and Orthogonal Transform for Bilingual Word Translation.” Pp. 1006–11 in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics.

Yamins, Daniel L. K., and James J. DiCarlo. 2016. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” *Nature Neuroscience* 19(3):356–65. doi: 10.1038/nn.4244.

Zäske, Romi, Marie-Christin Perlich, and Stefan R. Schweinberger. 2016. “To Hear or Not to Hear: Voice Processing under Visual Load.” *Attention, Perception, & Psychophysics* 78(5):1488–95. doi: 10.3758/s13414-016-1119-2.

Zhang, Yang, Yue Ding, Juan Huang, Wenjing Zhou, Zhipei Ling, Bo Hong, and Xiaoqin Wang. 2021. “Hierarchical Cortical Networks of ‘Voice Patches’ for Processing Voices in Human Brain.” *Proceedings of the National Academy of Sciences* 118(52):e2113887118. doi: 10.1073/pnas.2113887118.

Acknowledgements

We thank Bruno Nazarian for the design of an MRI-compatible button. We thank Jean-Luc Anton and Kepkee Loh for useful discussions. This work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 788240). This work was performed in the Center IRM-INT@CERIMED (UMR 7289, AMU-CNRS), platform member of France Life Imaging network (grant ANR-11-INBS-0 0 06). This work, carried out within the Institute of Convergence ILCB (ANR-16-CONV-0002), has benefited from support from the French government (*France 2030*), managed by the French National Agency for Research (ANR) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).