Separable Neural Representations of Sound Sources: Speaker Identity and Musical Timbre

Mattson Ogg[a], Dustin Moraczewski[a], Stefanie E. Kuchinsky[b,a], L. Robert Slevc[a]

[a] University of Maryland, College Park
[b] Walter Reed National Military Medical Center

This is a preprint of an article in *NeuroImage* – the final publication is available at:
https://doi.org/10.1016/j.neuroimage.2019.01.075

Correspondence should be sent to Mattson Ogg, Neuroscience and Cognitive Science Program, Department of Psychology, University of Maryland, Biology-Psychology Building, Room 3150 4094 Campus Drive, College Park, MD 20742 or via email at: mogg@umd.edu

Abstract

Human listeners can quickly and easily recognize different sound sources (objects and events) in their environment. Understanding how this impressive ability is accomplished can improve signal processing and machine intelligence applications along with assistive listening technologies. However, it is not clear how the brain represents the many sounds that humans can recognize (such as speech and music) at the level of individual sources, categories and acoustic features. To examine the cortical organization of these representations, we used patterns of fMRI responses to decode 1) four individual speakers and instruments from one another (separately, within each category), 2) the superordinate category labels associated with each stimulus (speech or instrument), and 3) a set of simple synthesized sounds that could be differentiated entirely on their acoustic features. Data were collected using an interleaved silent steady state sequence to increase the temporal signal-to-noise ratio, and mitigate issues with auditory stimulus presentation in fMRI. Largely separable clusters of voxels in the temporal lobes supported the decoding of individual speakers and instruments from other stimuli in the same category. Decoding the superordinate category of each sound was more accurate and involved a larger portion of the temporal lobes. However, these clusters all overlapped with areas that could decode simple, acoustically separable stimuli. Thus, individual sound sources from different sound categories are represented in separate regions of the temporal lobes that are situated within regions implicated in more general acoustic processes. These results bridge an important gap in our understanding of cortical representations of sounds and their acoustics.

Keywords: fMRI, Auditory Object, Auditory Perception, Speaker Identification, Speech, Timbre, Music, MVPA

1. Introduction

The identification of sound sources in the environment is a crucial function of auditory perception that is central to auditory scene analysis (Bizley and Cohen, 2013; Bregman, 1990), speech perception (Creel and Bregman, 2011) and music perception (McAdams and Giordano, 2009). In everyday life, listeners identify sounds quickly and easily. The ease with which one can accomplish such a task belies a difficult computational problem that the human auditory system must solve: how does the brain represent the multitude of sound sources a listener can identify? A better understanding of the neural processes supporting listeners' efficient object identification abilities could lead to improved signal processing and machine intelligence algorithms which can translate into better assistive therapies and devices to help relieve the difficulty that hearing loss poses for an aging population.

One way the human auditory system could represent behaviorally relevant sound sources, such as conspecific voices or musical instruments is through separate neural resources attuned to a particular class of sounds for efficient read-out by other cognitive or neural operations (e.g., Norman-Haignere et al., 2015). Alternatively, all types of sound sources might be represented via the same neural substrates (e.g., Zatorre et al., 2004), with little or no category-specific operations until later levels of analysis. However, previous work has primarily examined either individual sources within a single category (Bonte et al., 2014; Formisano et al., 2008; Hjortkjær et al., 2017), or broad category-level differences between large groups of sounds (Lee et al., 2015; Staeren et al., 2009). As a consequence, it remains unclear whether individual auditory objects or events from different categories are represented using the same or different neural substrates.

Functional magnetic resonance imaging (fMRI) studies have implicated the anterior superior temporal lobes in the processing of speaker identity (Belin and Zatorre, 2003; Chandrasekaran et al., 2011; Von Kriegstein et al., 2003). Furthermore, studies using multivariate pattern analysis (MVPA) have shown that patterns of fMRI activation in the superior temporal lobes can distinguish the individual speaker that a listener has just heard (Bonte et al., 2014; Formisano et al., 2008). In this way, MVPA is a useful tool for studying object perception because it permits the localization of neural information relevant to a particular stimulus or condition (Kriegeskorte et al., 2006). Additionally, MVPA can be more sensitive than univariate techniques because it takes advantage of the entire pattern of (often subtle) activation within a region (Haynes and Rees, 2006; Tong and Pratte, 2012, but see Akama et al., 2012; Jimura and Poldrack, 2012).

The superior temporal lobes and planum temporale have also been implicated in processing musical sound sources (Alluri et al., 2012; Menon et al., 2002; Halpern et al., 2004), which (like speakers) are particularly salient to human listeners. The musical quality that refers to different sound sources (instruments) is known as timbre (McAdams & Giordano, 2009). Similarly, studies of musical sounds in relation to diverse sets of speech and environmental sounds have identified music-specific responses in the anterior temporal lobes and planum polare (Angulo-Perkins et al., 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015; review: Ogg and Slevc, in press). However, such diverse stimulus sets necessarily involve greater acoustic variability, which can be represented in patterns of activation (Allen et al., 2017; Giordano et al., 2012) that a classifier could exploit. Therefore, when using diverse stimulus sets, steps must be taken to account for these acoustic processes either by controlling the features of the stimuli themselves (Lee et al., 2015; Staeren et al., 2009), or by incorporating the acoustic

variability into the analyses (Hjortkjær et al., 2017; Giordano et al., 2014). Additionally, these studies involving musical timbre were primarily based on univariate activation. MVPA would permit the examination of individual object (Hjortkjær et al., 2017) and category (Lee et al., 2015; Staeren et al., 2009) representations.

In sum, results from both activation and decoding studies suggest a common reliance on the anterior temporal lobes as well as planum temporale for the processing of various categories of auditory objects and events. However, this work has involved different levels of granularity regarding the comparisons, sound categories, and acoustic features examined. Thus, the present study used a searchlight decoding approach to directly compare the decoding of natural recordings of individual auditory objects or events from different categories of sound that are highly relevant for human listeners: speakers and instruments. Our stimuli controlled for many low-level acoustic parameters, but were allowed to naturally vary along dimensions known to facilitate robust and ecologically valid object recognition. Additionally, we obtained behavioral identification measures during an initial familiarization phase that ensured participants could accurately perceive and differentiate these individual sound sources, which might also provide context for the neural results. Finally, our use of an interleaved silent steady state acquisition (ISSS) MRI sequence (Schwarzbauer et al., 2006) allowed for optimal presentation of the sound stimuli within gaps between the loud scanner volume acquisition noise (a persistent issue with auditory studies in MRI; Peelle, 2014), while maximizing the temporal signal-to-noise ratio in our data by averaging over multiple scans (Murphy et al., 2007).

If the perception of individual instruments can be decoded from patterns of fMRI activation similarly to individual speakers (Bonte et al., 2014; Formisano et al., 2008), it would suggest that the brain represents individual auditory objects (instruments or speakers) in a

general manner (based on common neural substrates). Alternatively, if they are decoded in different areas of the brain, it would suggest there are separate neural resources devoted to individual sound sources from different categories. To account for how our classification results might reflect neural representations of the low-level acoustic features of these sounds, we also identified regions associated with basic auditory processes using simple synthesized auditory stimuli. This approach allowed us to examine object, category and acoustic level sound representations in the brain.

2. Materials and Methods

2.1 Participants

We recruited 18 participants (9 female, age $M = 22.7$, $SD = 3.8$) from the University of Maryland community. All participants were right handed, fluent English speakers, with normal hearing, and no neurological disorders based on self report. The presence or absence of musical training was not exclusionary, and most participants had some degree of musical training ($M =$ 5.1 years, $SD = 3.6$), which is typical of a university-based participant population (Corrigall et al., 2013). Participants gave informed consent to participate and were compensated monetarily. The procedures in this study were approved by the University of Maryland Institutional Review Board.

2.2 Stimuli

2.2.1 Main Study: We presented participants with eight different sound sources. Four of these sources were different speakers of American English (vowels spoken by two males and two females) who were not known to the participants prior to study participation. The other four

sound sources were different musical instruments (bass clarinet, bowed cello, marimba, and trombone) chosen to broadly represent the different families of orchestral instruments. We used three different tokens for each sound source (i.e., each speaker or instrument) which required the classifier to generalize across unique tokens to decode a given sound source. Example spectrograms are plotted in Figure 1, along with a schematic of the experimental design.

Note that our focus in constructing this stimulus set was not to comprehensively control for all the low-level acoustic features that might differ between or among music and speech sound sources. Instead, our view was that natural and important acoustic differences exist in terms of spectral or temporal envelopes (among instrument timbres) and fundamental frequency (among speakers) that support object recognition in the auditory system. These stimuli were selected to represent such qualities and to therefore facilitate robust and ecologically valid object recognition for the listener.

For each of the four speakers, we created tokens that consisted of three different vowels which were extracted from longer consonant-vowel-consonant utterances. Vowels (rather than consonants) were used because they are particularly useful for listeners when identifying different speakers (Owren and Cardillo, 2006) and for decoding speaker-related information from neural responses (Formisano et al., 2008; Khalighinejad et al., 2017). Speakers read a randomized list of utterances that crossed the consonants /h/, /b/, and /g/ with the vowels /ɑ/, /i/ and /u/ and ended with /d/ (resulting in the utterance “heed,” for example). From these, one instance of each vowel (/ɑ/, /i/ and /u/) was extracted for each speaker. Speakers were instructed to keep the pitch of their voices as consistent as possible but to otherwise speak normally. Recording equipment and conditions for the speech stimuli were similar to previous work (Ogg et al., 2017).

For each of the four instruments, we used tokens consisting of three different notes (obtained from University of Iowa Musical Instrument Samples Database; 1997), which served two methodological functions. First, as with the different speakers' vowels, the acoustic variability from the different notes required the classifier to generalize across neural responses associated with a particular stimulus token, resulting in a stronger test of sound source decoding. Second, this allowed us to match the fundamental frequencies of the instruments to the speech utterances, which could otherwise provide a trivial cue that the classifier might use. To accomplish this, we calculated the median fundamental frequency of each speech token using the YIN algorithm (de Cheveigné and Kawahara, 2002) and identified the closest musical note. We then randomly assigned these notes (3 x $A\flat_2$, 1 x $B\flat_2$, 2 x C_3, 3 x $A\flat_3$, 2 x B_3, 1 x C_4; which correspond to approximately 103.83, 116.54, 130.81, 207.65, 246.94, and 261.63 Hz, respectively) to the musical instruments. The lower fundamental frequencies of the male speakers required the use of lower register musical instruments. Note that assignment was constrained such that no instrument played the same note twice and each instrument played at least one note corresponding to a male and a female utterance.

Using these different pitches means that the organization of fundamental frequency is preserved among the stimuli where this cue is a naturally important dimension, such as for distinguishing among speakers. At the same time this control obviates fundamental frequency as a useful cue in other comparisons where this feature varies more arbitrary with respect to individual sound sources, such as for distinguishing instruments, or distinguishing between instruments and speakers. In the former case, listeners and the classifier can still use fundamental frequency as a cue to discriminate stimuli, whereas in the latter cases it must be ignored.

Thus, the eight sound sources we aimed to decode were conveyed to the participants via 24 unique stimuli: 12 instrument tokens comprising 3 notes from 4 instruments, and 12 speech tokens comprising 3 vowels from 4 speakers. We also included a cat vocalization sound as a catch trial to ensure that participants remained attentive during the scan.

Each stimulus was edited to its most identifiable 230 ms section. We selected this duration for all stimuli to prevent duration from being a distinguishable cue between stimuli and because this was the longest duration of the vowels that minimized any notable coarticulation from the adjacent consonants. Vowels were manually extracted from the speech utterances, then visually and aurally inspected to confirm they contained consistent periodic waveforms and minimal change due to coarticulation (Figure 1). Because sound onset dynamics are integral to the perception of musical timbre (McAdams and Giordano, 2009), instrument stimuli (and the cat stimulus) were edited to a 230 ms window beginning at each token's onset, defined here as 5 ms prior to when the sound first exceeded 10% of its maximum absolute amplitude. We then applied 20 ms onset and offset cosine ramps, and RMS level normalization to all tokens. Finally, stimuli were filtered to achieve a flat frequency response (+/– 2dB between) between 50 Hz and 6000 Hz, where the frequency response of the Sensimetrics playback system (Malden, MA) and Etymotic 3A insert earphones (Etymotic Research, Elk Grove Village, Illinois) falls off. Stimuli were played back in the scanner at approximately 72 dBA SPL, although the level was allowed to be increased per participant request.

2.2.2 Auditory Localizer: We also presented participants with simple, synthesized stimuli that have been used in previous studies to identify early auditory processes (Figure 2). These were non-overlapping patterns of sinetones centered around 200, 400, 800, 1600, 3200, and 6400 Hz that resembled "auditory checkerboards" (as in Humphries et al., 2010). Each checkerboard

consisted of fifteen 80 ms bursts of two (at 0.9 or 1.11 times the center frequency) or three (at 0.83, 1 or 1.23 times the center frequency) tones that were not harmonically related with 20 ms onset and offset cosine ramps presented at a rate of 10 Hz. The order of the two or three tone bursts within each trial was randomized for a total of five unique patterns within each frequency range. We also presented participants with five different instances of 8 Hz amplitude modulated gaussian white noise (similar to Okada et al., 2010). Each localizer stimulus lasted 1.5 seconds and was RMS-normalized to the same level as the speech and instrument stimuli and the same filters were applied. The checkerboard and noise stimuli were presented together in a randomized order within each localizer block (thus, the noise stimuli and each checker board frequency range made up $1/7^{th}$ of the localizer stimuli).

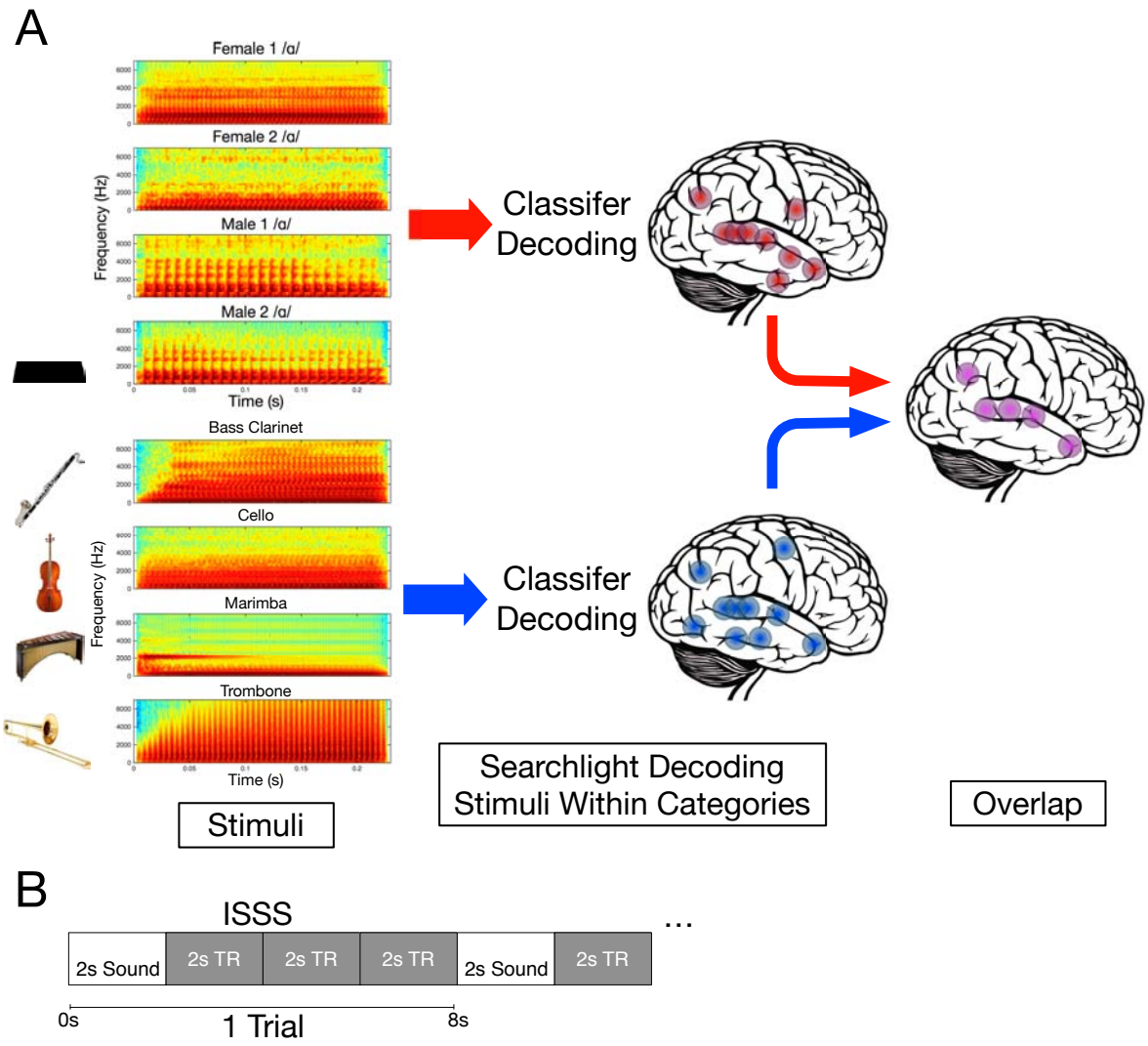

Figure 1. Schematic of the experimental stimuli and design. A: picture and spectrogram depictions of the sound sources within each category (left). The vowels /i/ and /u/ from each speaker were also presented (not pictured), as well as two other notes for each instrument for a total of three vowels from each of four speakers and three notes from each of four instruments. The sound sources (speakers and instruments) were each decoded from one another via a series of searchlights whose center voxels iterated through all the voxel positions in the brain. The speaker and instrument decoding results were then compared with one another to determine which voxels or searchlights, if any, were able to significantly decode stimuli in both sound categories. Colored circles are illustrative and do not depict actual data or searchlights. B:

diagram of stimulus presentation and TR acquisitions in the interleaved silent steady state scanning sequence for a given trial.

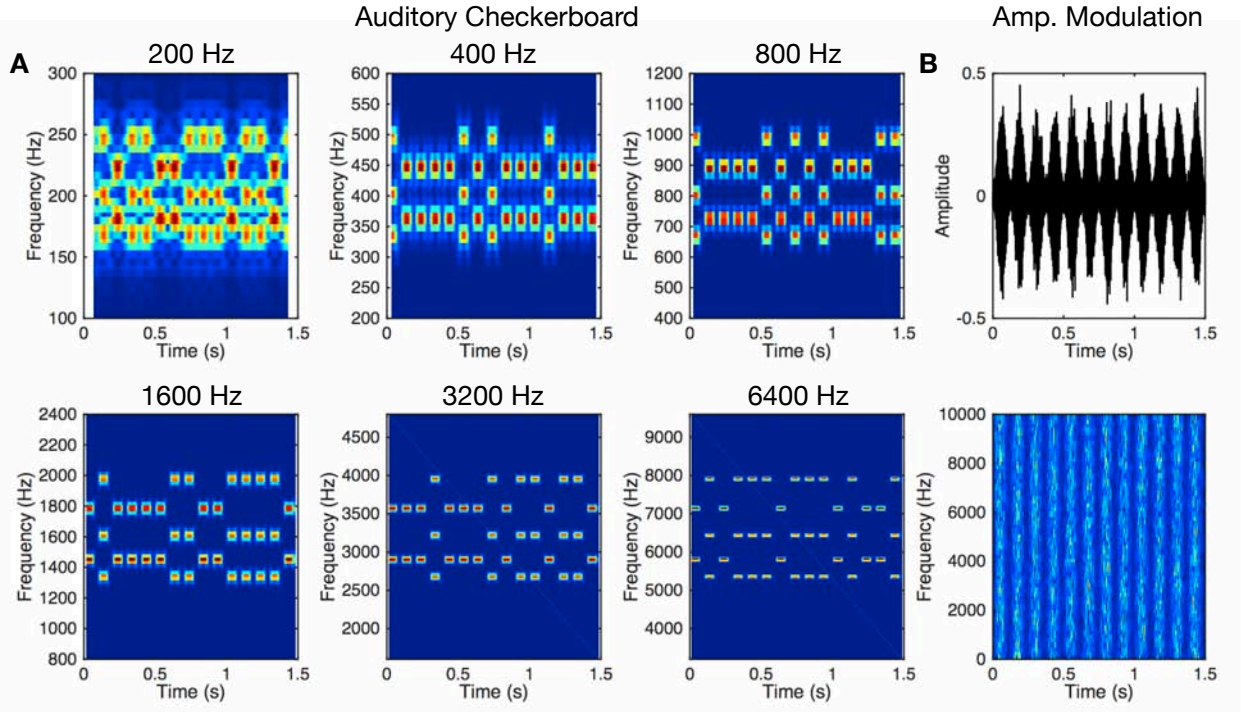

Figure 2. Examples of A: the auditory checkerboard and B: amplitude modulated noise stimuli used in the auditory localizer. Tone frequencies are plotted on a linear frequency axis: at 0.9 or 1.11 times the center frequency or at 0.83, 1 or 1.23 times the center frequency. The center frequencies of each checkerboard are indicated above each plot. Spectrogram window and overlap parameters were adjusted across plots in A to more clearly indicate the separation of the tones in different frequency ranges.

2.3 Familiarization and Screening

Participants completed a familiarization and screening session no more than eight days prior to the MRI session to introduce them to the stimuli and to ensure they could accurately identify the speaker and instrument sound sources. Accurate identification would 1) Confirm a comparable level of difficulty in perceiving these sound sources; 2) Suggest that participants' auditory responses to the stimuli were distinct from one another at some level of processing which the classifier might potentially discern and; 3) Provide a larger context for our interpretation of the decoding results. In the familiarization and screening phase, participants first freely listened to each of the 25 stimulus sound files (24 critical tokens plus the cat vocalization) as many times as they liked. Next, participants completed six blocks of trials in the familiarization and screening paradigm (administered via PsychoPy, version 1.83.4; Peirce, 2007). Each trial presented the participant with three repetitions of a single stimulus, with 700 ms inter-stimulus-intervals. In each block, stimuli were presented in a random order.

In the first block of six blocks of trials, participants listened passively to the stimuli while the name and a picture of each sound source was displayed on screen. In the second block of trials, stimuli were presented one at a time and participants were required to indicate the sound source via button press (closed set with nine response options on screen: each of the four instruments and speakers plus the cat). In the second block, participants were given feedback to indicate if they were correct or, if not, to indicate the correct sound source. The third block of trials was designated as the test block and was identical to the second block except there was no feedback.

The final three blocks of trials were identical to the first three, but instead of the original sound files, participants were presented with recordings of each stimulus made from the earpiece of the MRI scanner sound delivery system (while the scanner was not running). This was done to

ensure that participants could identify the stimuli despite any changes that might be imparted by the scanner playback apparatus. Recordings of the stimuli were made via a Larson Davis (Larson Davis, Depew, NY) 2cc coupler and 1/2in condenser microphone, 824S/PRM902 preamplifier and sound level meter, and an SD702 digital recorder (Sound Devices, Reedsburg, WI).

Screening performance was assessed using the participant's accuracy on the sixth block (i.e., second testing block, based on recordings from the scanner's playback system). Participants were required to achieve a criterion of at least 22 out of 25 (88%) correct in order to proceed to the scan. If a participant did not achieve this high accuracy on their first try, they were permitted to listen back to the stimuli at their own pace or repeat any portion of the screening before trying the critical testing block again. Three participants required a second try on the testing block and one required a third, two additional participants did not meet this criterion and thus did not undergo a scan.

2.4 Data Acquisition

Scanning was performed on a Siemens 3T MAGNETOM Tim Trio MRI (Siemens AG, Munich, Germany) with a 32-channel head coil. A high resolution, T1-weighted structural scan was acquired via an MPRAGE sequence (192 contiguous sagittal slices; 0.9×0.9×0.9 mm voxel size; TR = 1900 ms; TE = 2.32 ms; flip angle = 9°). Functional scanning was then performed via an interleaved silent steady state acquisition sequence (ISSS; Schwarzbauer et al., 2006). Compared to continuous or fully sparse sampling acquisition approaches, this sequence maximizes the temporal signal-to-noise ratio by averaging the BOLD response across multiple functional volumes (Murphy et al., 2007), while still providing a quiet period during which the stimuli could be played without interference from the loud acquisition sequence (Peelle, 2014).

Thus, each trial began with a 2-second epoch without the scanner acquisition noise, in which pulses maintain the magnetic field and allow for stimulus presentation in relatively quiet conditions. This was followed by the acquisition of three functional volumes (voxel size = 3x3x3-mm; 32 slices; TR = 2000 ms; TE = 30 ms; flip angle = 78 degrees; gap = .75-mm) also lasting 2 seconds each for a total trial length of 8 seconds. On screen instructions, participant responses, stimulus presentation and timing were controlled via a PC running PsychoPy (version 1.83.4; Peirce, 2007). Visual instructions were presented on a screen at the back of the bore projected on a head-coil-mounted mirror.

After the structural scan, participants underwent six runs of 60 trials each in the main study (for a total of 180 volumes per run). The participant's task in the main study was to push a button every time they heard the cat vocalization stimulus (there was no task in the localizer runs). During the silent epoch (white portion of Figure 1B), a fixation cross was displayed and a given 230 ms token was presented three times with a 405 ms ISI and a 250 ms silent buffer on either end of the two second presentation window. This was followed by a visual prompt to press the button if that trial contained the cat sound, which remained onscreen for the rest of the trial's six second duration. The visual prompt was the same on every trial. Catch trials where the cat stimulus was presented were interspersed with the speech and instrument tokens and were also presented three times per trial (silent epoch) in the same manner as the other tokens (Figure 1B). Each of the unique 24 stimuli was presented twice per run for a total of six instances of each speaker and instrument sound source. Each run also contained eight baseline trials (one at the beginning and end of each run) in which no auditory stimulus was presented, as well as four catch (cat) trials. Within each run, stimuli were presented in pseudorandomized order,

constrained such that no speaker or instrument was presented twice in a row. The catch trials and responses were excluded from the decoding analyses.

After the six runs of the main study, participants underwent three runs of the auditory localizer. The functional parameters and protocol for the auditory localizer were the same as above except for the following differences: Five exemplars of each of the seven synthesized stimuli (amplitude modulated noise, or checkerboards centered on each of the six frequencies) were played per run along with seven baseline trials in which no stimulus was presented (one at the beginning and end of the run) for a total of 42 trials per run (126 volumes per run). Stimuli were again pseudo randomized such that no frequency range (or noise stimulus) was presented twice in a row. Localizer stimuli were played continuously for 1.5 seconds in the middle of the two-second silent epoch. Participants were instructed to listen carefully to the stimuli and were not required to perform any task during the localizer.

2.5 Preprocessing

All preprocessing steps were carried out in AFNI version 18.0.18 (Cox, 1996). The functional volumes were first re-aligned to the first volume of the participant's first run, co-registered with the participant's own structural image, and spatially normalized using affine transformation to a standard MNI template. During the co-registration step, six motion parameters (3 translation and 3 rotation) and their derivatives were recorded for inclusion in the regression analyses. All further analyses used the final two of the three functional ISSS volumes collected after stimulus presentation. This was done for two primary reasons: 1) Given that the canonical hemodynamic response peaks 4 to 6 seconds after a stimulus, we focused our analysis on the final two of the three functional volumes collected after stimulus presentation; and 2)

Inspection of the data revealed that the steady-state longitudinal magnetization had been insufficiently maintained during the silent epoch of the ISSS sequence, resulting in a T1-signal decay for only the first of the three acquisitions (see Supplementary Figure 1). To focus on the second and third volumes of interest, we removed the first ISSS volume from the time series and concatenated the remaining volumes for decoding and activation analyses. In addition, we also removed motion parameters and derivatives that corresponded to the removed functional volumes. The resulting BOLD timeseries was spatially-normalized and then entered into the following preprocessing steps for decoding and activation analyses, respectively.

For the decoding analyses (both the main study and localizer), motion parameters and their derivatives (for the second and third volumes), along with terms for linear and quadratic low-frequency drift for each run were removed using a linear regression. The residuals were then masked using an individually-defined mask to include only voxels within the brain. The somewhat small field of view employed along with the angle that captured the temporal lobes meant that in most cases we could not capture a participant's entire cortex. For these participants, we prioritized capturing the temporal lobes which reduced coverage for the dorsal and anterior-most portions of the parietal and frontal lobes, respectively. We report results only for voxels where signal was present in all participants. Each voxel's response was then averaged across the two scans for each trial and these trial-level voxel responses were used as feature vectors for training and testing the classifier.

For univariate activation analysis, re-alignment, co-registration, and spatial normalization was performed using identical steps as the above decoding preprocessing. We then normalized the BOLD signal to a mean of 100 using the voxelwise mean intensity, and spatially smoothed voxels within individually-defined brain masks using a 5-mm FWHM gaussian kernel. The first-

level general linear model (GLM) analysis used a finite impulse response (FIR) model, in which we created a design matrix that included impulse response regressors for the two volumes that followed each sound condition (speech or instrument) and a nuisance impulse regressor for the catch-trials. In addition, we included the de-meaned motion parameters and their derivatives (for the second and third volumes on each trial) as nuisance regressors. A post hoc speech > instrument contrast was performed, which was then submitted to a second-level analyses across participants. Group analysis was performed using a mixed-effect multilevel model (3dMEMA in AFNI), which weights effect estimates by their variance (Chen et al., 2012). The results were corrected for multiple comparisons using 1000 Monte Carlo simulations to assess statistically significant clusters that could arise by chance while also accounting for the spatial autocorrelation structure of the data (using 3dClustsim with the ACF option in AFNI; Cox et al., 2017).

2.6 Stimulus Decoding

Classification analyses were performed at the individual participant level with a searchlight algorithm implemented in CoSMoMVPA (Oosterhof et al, 2016) in MATLAB (2014b, MathWorks, Natick, United States). Separate searchlights were run to decode the instrument timbres from one another, the speakers from one another (both 4-way classification, chance = 25% separated by subsetting the trials by category; see Figure 1a), and whether a given sound was a speech or instrument sound (2-way classification, chance = 50%, no subsetting aside from removing baseline and catch trials). To ensure that any potential differences in classification performance from the 4-way, within-category analyses were due to better representations of the categories themselves rather than simply a greater amount of training data

(30 training cases per within-category target compared to 120 for between categories), we randomly selected a comparable subset of trials per run for each participant before conducting the between-category searchlight (thus yielding 30 trials per target: ”instrument” or “speech”, constrained such that each speaker or instrument under each category label was present at least once per run).

Baseline and catch trials were removed from the lists of trials prior to decoding. Each searchlight used a linear discriminant classifier with a 3-voxel radius (average size = 113.4 voxels), such that each voxel in the brain served as the center location of a searchlight. We used a 6-fold, leave-one-run-out cross-validation scheme to train and test the classifier. This process, resulted in a map of classification accuracies averaged over cross-validation folds.

The same procedure was used to decode the auditory localizer conditions. The labels the classifier decoded for the localizer were the center frequencies of the checkerboards or “noise.” This searchlight analysis was conducted with 3-fold, leave-one-run-out cross-validation using the trials in the three localizer runs (compared to the six runs available for the main study), again excluding silent baseline trials. We obtained the same pattern of results and conclusions when the noise stimulus was excluded from the localizer decoding analysis.

Statistical significance for the decoding results was assessed at the group level via permutation testing and threshold free cluster enhancement (Smith and Nichols, 2009; Stelzer, Chen, and Turner, 2013). For each searchlight analysis, we generated 100 permuted accuracy maps for each participant by shuffling the labels of the stimuli for training and testing within each run and then repeating the decoding algorithm above. These permuted participant-level maps were then included among 10,000 bootstrapped sign permutation testing iterations and threshold free cluster enhancement as implemented in CoSMoMVPA (Oosterhof et al., 2016).

For each searchlight, this yielded *z*-statistics that we assessed as one-tailed tests against chance performance at a criteria of $p < 0.05$ ($z > 1.6449$) corrected for multiple comparisons. All results are visualized using MRIcroGL (http://www.mccauslandcenter.sc.edu/mricrogl/).

3. Results

Participants were able to accurately identify these individual sound sources in a behavioral identification task conducted at screening (*M* identification accuracy = 93%, *SD* = 3.6%), indicating that these different sound sources were clearly distinguishable to the participants. Moreover, performance on the identification task did not differ between the speech and instrument stimuli (paired sample t-test: $t_{(17)} = 1.22$, $p = 0.24$; speech $M = 95\%$, $SD = 6.5\%$; instrument $M = 91\%$, $SD = 8.3\%$), indicating a comparable level of difficulty for discerning among these sets of sound sources. To ensure participants maintained attention in the scanner, their task was to push a button every time a trial consisted of a cat vocalization. Participants were very accurate on this vigilance task (*M* accuracy = 96.8%, *SD* = 8.2%) indicating that they were attentive to the sounds throughout the scan. The cat trials were not analyzed further.

To understand where individual sound sources, sound categories, and basic acoustic features were represented in the brain, we first report a series of searchlight decoding analyses (in that order). Next, we describe how the decoding of individual sound sources and categories might be accounted for by low-level acoustic representations. This is followed by a univariate analysis of overall neural activation for the speech and instrument tokens. Finally, we examine the potential influence that individual factors might have had on these results such as musical training or behavioral identification at screening.

3.1 Mapping Representations of Individual Speakers and Instruments: Within-Category Searchlight Analyses

We first investigated whether the unique sound sources within each category (speakers or instruments) could be decoded from one another, and whether this involved overlapping or separate neural substrates among the different categories. Thus, we had participants listen to multiple examples of four speakers and four instruments, and then performed a searchlight decoding analysis of the stimuli within each category (Figure 3A and Table 1).

The instruments that participants heard could be decoded from one another in 428 voxel searchlight locations (all corrected $p < 0.05$) in portions of the bilateral temporal lobes. These regions spanned an area posterior to primary auditory cortex that extended anteriorly along superior temporal gyrus and to the right inferior frontal gyrus, including Heschl's gyrus in the right, but not the left hemisphere. This pattern was more lateral in the left hemisphere and extended more anteriorly in the right hemisphere. In contrast, the individual speakers that participants heard could be decoded from one another in a circumscribed cluster of 18 voxel searchlight locations (all corrected $p < 0.05$), all in the left hemisphere near primary auditory cortex, encompassing parts of the transverse temporal gyrus and insula. Interestingly, only one voxel (searchlight location) was found to be able to significantly decode stimuli from both categories, located in the left superior temporal gyrus. Group averaged accuracy rates among the voxels in significant clusters were 3.5% (SE = 0.3%) better than chance performance for the instrument searchlight, and 4.2% (SE = 0.5%) better than chance performance for the speaker searchlight (where chance for both = 25%), although the maximum decoding accuracy that was achieved for individual participants was much higher (Figure 3). These results suggest that individual timbres and speakers can be decoded from one another using this searchlight method,

but that this decoding appears to involve largely separable neural resources between speech and instrument sounds.

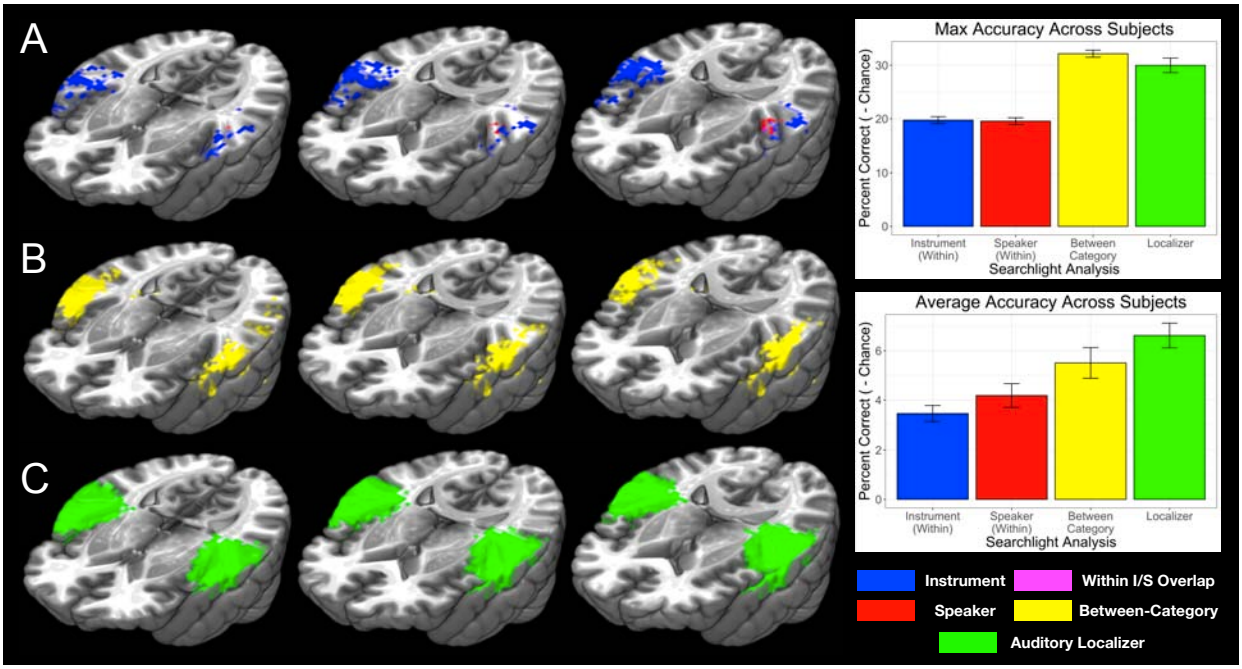

Figure 3. Searchlight decoding results. The A: separate within-category (instrument vs. instrument and speaker vs. speaker), B: between-category (instrument vs. speech), and C: auditory localizer searchlight analyses overlaid onto the same template brain. All clusters were significant ($p < 0.05$) based on 10,000 bootstrapped sign permutation tests against chance (one-tailed; using 100 null permutation maps for each participant) and threshold free cluster enhancement. Within-instrument decoding is depicted in blue, within-speaker decoding is depicted in red, overlap between these is depicted in violet, between-category decoding is depicted in yellow, and the auditory localizer is depicted in green. Maximum participant level accuracy rates in each analysis (minus chance) as well as average participant level accuracy rates within the significant group-defined clusters (minus chance) are summarized in the plots on the right.

3.2 Mapping Representations of Sound Categories: Between-Category Searchlight Analysis for Speech vs. Instrument

Following up on the previous analysis that examined the classification of individual instruments or speakers from one another, we also investigated whether patterns of activation associated with each stimulus' superordinate sound category (speech or instrument) are linearly separable from one another. Thus, we ran an additional searchlight analysis where a classifier was trained and tested on the category labels "speech" and "instrument" associated with each stimulus.

These results (Figure 3B, Table 1) revealed significant bilateral decoding extending from the posterior parts of left and right superior and middle temporal gyrus to the anterior superior temporal gyrus among 1032 voxel searchlight locations (all corrected $p < 0.05$). Group averaged accuracy rates among voxels in these significant clusters were 5.5% ($SE = 0.6\%$) better than chance performance (chance = 50%), although again the maximum decoding accuracy that was achieved for individual participants was much higher (Figure 3). Between-category decoding accuracy was much higher than the within-category decoding accuracy (see Figure 3) and involved a more extensive portion of cortex (1032 significant voxels compared to 428 or 18, see Table 1), suggesting that the neural representations of these superordinate sound categories were more linearly distinguishable than those of the individual stimuli within each category. Interestingly, the between-category decoding map implicated voxels that only partially overlapped with the two within-category decoding maps: three voxels overlapped with the speaker decoding map (16.67% of the speaker voxels) and 174 voxels overlapped with the instrument decoding map (40.65% of the instrument voxels).

Table 1. Searchlight Decoding Analysis Clusters

Description	*Z*	*Voxels (n)*	*MNI Coordinates (X, Y, Z)*
Within-Instrument			
L. Superior Temporal Gyrus	2.61	125	–58, –31, 5
R. Superior Temporal Gyrus	3.43	301	61, –16, 5
Within-Speech			
L. Transverse Temporal Gyrus	2.05	18	–45, –22, 12
Within-Category Overlap			
L. Superior Temporal Gyrus		1	–48, –16, 8
Between-Category			
L. Superior Temporal Gyrus	3.12	460	–61, –13, 6
R. Superior Temporal Gyrus	3.72	528	58, –12, 2
Auditory Localizer			
L. Superior Temporal Gyrus (Center)	3.72	1390	–50, –23, 7
R. Superior Temporal Gyrus (Center)	3.72	1457	55, –19, 6

Only clusters of at least 10 voxels are displayed with the exception of within-category overlap.

All clusters were significant ($p < 0.05$) based on 10,000 bootstrapped sign permutation tests against chance (one-tailed; using 100 null permutation maps for each participant) and threshold free cluster enhancement.

Maximum *z*-value returned by the bootstrap algorithm for 10,000 iterations is 3.72.

All listed *z*-values pertain to statistical peaks unless otherwise noted as a center coordinate (in the case of large clusters of contiguous maximum *z*-values).

3.3 The Influence of Low-Level Auditory Features: Auditory Localizer Decoding

The within- and between-category searchlight results suggest neural response patterns that distinguish different sound sources and categories from one another, but thus far it is not entirely clear what underlying neural processes these decoding results reflect. We selected portions of the speech and instrument tokens that the literature indicated would be most identifiable: the vowels of utterances of speakers (Owren and Cardillo, 2006), and the onsets of instrument notes (McAdams and Giordano, 2009). Moreover, we allowed the acoustic features that support sound identification within these categories to vary naturally: spectral and temporal envelopes among instruments and fundamental frequency among speakers. This could precipitate different responses in the regions of auditory cortex responsible for representing these acoustic properties, which in turn might have been exploited by the classifier.

To determine the extent to which our findings thus far might rely on such low-level acoustic processing mechanisms, we decoded neural responses to simple synthesized tone patterns centered at different frequencies, as well as amplitude modulated noise, all of which have been used in previous studies to localize primary auditory functions (Humphries et al., 2010; Okada et al., 2010). The logic of this approach is that if the classification of speakers or instruments in a particular region capitalized on neural representations of low-level acoustic feature differences that might exist among these natural sounds, then those same regions should also be sensitive to (i.e., represent) the different frequency ranges of the sinetone checkerboards and amplitude modulated noise used in the localizer. This was quantified by examining any potential overlap between the maps of regions that supported the significant decoding of natural sounds and maps of regions that the supported significant decoding of the localizer stimuli.

The auditory localizer revealed large clusters of significant decoding in 2847 voxel searchlight locations (all corrected $p < 0.05$) in the temporal lobes centered around primary auditory cortex in both hemispheres (Figure 3C). This activation extended from the insula and middle temporal gyrus anteriorly along superior temporal gyrus. The group averaged accuracy among voxels in these significant clusters was 6.6% ($SE = 0.5\%$) better than chance performance (chance = 14.29%) although again individual maximum decoding accuracies among participants was much higher (Figure 3).

The regions identified in the within-category searchlights were contained almost entirely within the set of voxels that significantly decoded the localizer stimuli (Figure 4). The speaker decoding map overlapped 100% with the localizer map, while the instrument decoding map overlapped 93.2% with the localizer map. This suggests that the within-category decoding was based largely on neural responses that pertained to the processing of low-level acoustic differences among these stimuli (but see Peretz, Vuvan, Lagrois, & Armony, 2015 for a discussion of the difficulty of interpreting overlap in fMRI). The only exception to this was a small cluster of 10 voxels in the right inferior frontal gyrus that was involved in decoding instruments but not involved in the decoding the localizer stimuli (MNI coordinate of cluster center: x = 59, y = 16, z = 6). Interestingly, the between-category searchlight, which identified significant voxels in the posterior and anterior temporal lobes, only overlapped 71% with the auditory localizer searchlight. This suggests that some other processes were responsible for between-category decoding in addition to low-level acoustic features.

Thus, the different decoding analyses were successful overall, however accuracy rates varied across tests (Figure 3). To quantify these differences, we ran a repeated measures ANOVA on the (logit transformed) maximum accuracy rates among participants in each

searchlight analysis. This revealed that the maximum decoding rates varied significantly across decoding analyses ($F_{(3,51)} = 67.61$, $p < 0.001$, $\eta_\rho^2 = 0.80$). Bonferroni-corrected post-hoc t-tests revealed that maximum accuracy rates for each test significantly differed from one another (all corrected $p < 0.001$) except for between the two within-category analyses and between the localizer analysis and the between-category analysis. This suggests that the neural representations of these different categories and different acoustic features were more reliably distinguishable than the neural representations of the individual stimuli within each category. Similar conclusions were reached based on ANOVAs of average participant-level accuracy rates within the significant group-defined clusters (Figure 3).

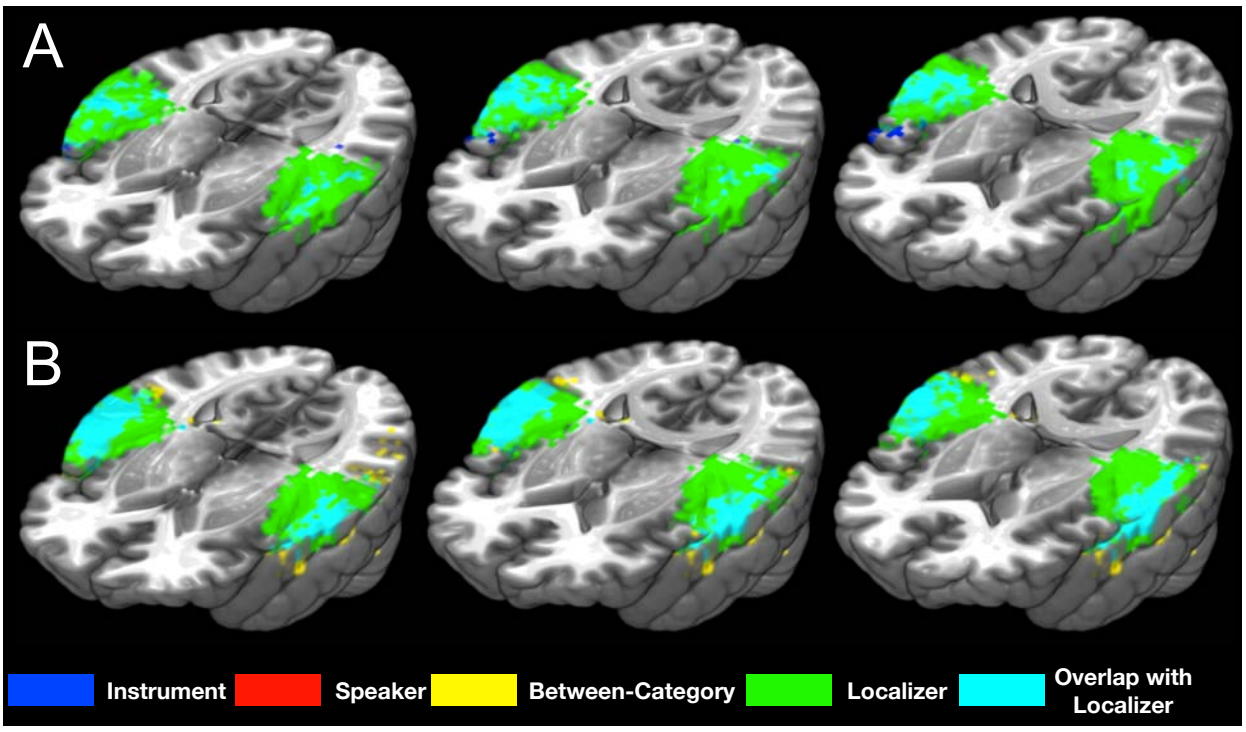

Figure 4. Overlap between the results of the auditory localizer searchlight analysis and A: the within category or B: between category decoding analyses overlaid onto a template brain. Colors denote the decoding analysis that the maps correspond to.

3.4 Univariate Activation for Speech and Instrument Stimuli

Our study design also allowed us to examine any potential differences in activation (rather than classification) between speech and instrument sounds. Thus, we compared the regions activated by the speech and instrument stimuli (averaging over the responses to stimuli within each category) in a subtractive analysis of univariate fMRI activation between these two categories (Figure 5, Table 2). This revealed a region that exhibited a stronger response to instrument stimuli around primary auditory cortex in the right hemisphere, along with two bilateral regions that expressed a preference for the speech stimuli located more ventrally and laterally in the temporal lobes compared to the instrument sensitive region (t-tests on beta estimates, uncorrected $p < 0.005$, with a 22 voxel cluster threshold, *FWE* $p < .05$, adjusted for autocorrelation among neighboring voxels).

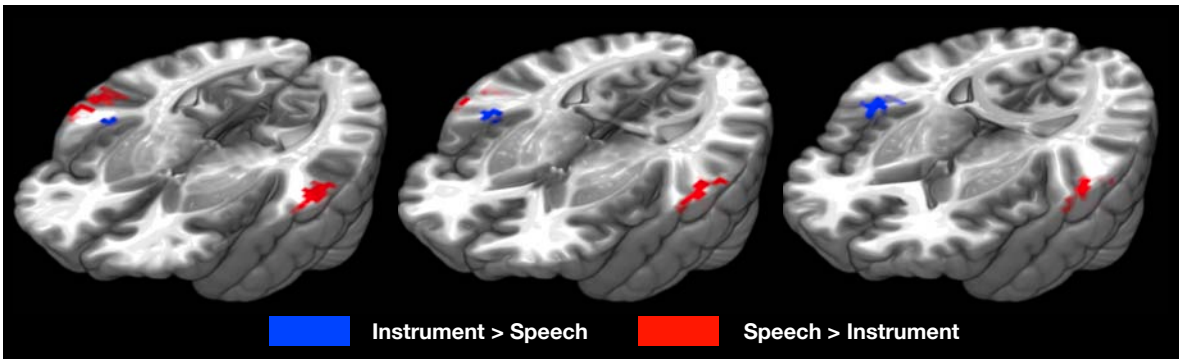

Figure 5. Results of the subtractive, univariate analysis of speech and instrument responses (speech > instrument). Voxels that exhibited stronger responses to speech sounds are plotted in red (positive t-values) and voxels that exhibited stronger responses to instrument sounds are plotted in blue (negative t-values), $p < 0.005$, 22 voxel cluster threshold, *FWE* $p < .05$, adjusted for autocorrelation among neighboring voxels.

Table 2 Peak Univariate Activation for the Contrast Speech > Instrument

Description	*t*	*Voxels (n)*	*MNI Coordinates (X, Y, Z)*
Instrument			
R. Superior Temporal Gyrus	–6.59	39	49,–22,9
Speech			
L. Middle Temporal Gyrus	5.37	57	–64,–21,–5
R. Middle Temporal Gyrus	5.42	53	69,–12,–4

All clusters significant at uncorrected $p < 0.005$, with a 22 voxel cluster threshold, *FWE* $p < .05$, adjusted for autocorrelation among neighboring voxels.

3.5 The Influence of Musical Training and Screening Task Performance

Comparable identification rates were observed among the musical instruments and speakers during the screening phase, which indicates a similar level of difficulty for listeners in perceiving both sets of sound sources (see section 3.1). However, our participant population did vary with respect to musical experience. Therefore, it is possible that participants with more musical training were more familiar with the instrument stimuli at baseline. Musical training has also been linked to general auditory perception advantages that could have influenced these results (Chartrand and Belin, 2006; Zendel and Alain, 2012). To examine the possible influence of musical training on familiarization and screening performance we calculated Pearson's product moment correlations between the participant's musical training subscale of the Goldsmith's Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014) and their overall task accuracy at screening as well as their accuracy on the subset of sounds within each stimulus category. These analyses were conducted both for initial task performance and when

participants reached the high criterion needed to proceed to the fMRI session (four participants needed to repeat portions of the screening task to reach criterion. See section 2.3).

This analysis revealed that musical training did not correlate with identification performance overall at screening, but we did observe a significant positive correlation between musical training and initial identification performance on musical instrument sounds ($r = 0.49$, $p = 0.04$). However, this relationship was not observed when participants eventually reached criterion ($p = 0.19$), suggesting that, despite some initial differences in familiarity with the instrument stimuli, all participants were able to achieve very accurate identification performance before scanning. No other significant correlations between behavioral performance and musical training were observed.

Next we examined whether the accuracy with which we were able to decode any of the sound representations from participants' fMRI responses (suggesting more robust neural representations for those sounds) or the strength of the univariate results might have related to musical training or screening performance. A significant positive correlation was observed only between the participants' maximum speech decoding accuracy rates and their scores on the musical training subscale of the Gold-MSI ($r = 0.56$, $p = 0.02$). No significant correlation with musical training was observed for musical instrument decoding or any other decoding outcome. We also re-examined the univariate analysis with musical training statistically controlled as a covariate, but this did not change the pattern of results or conclusions regarding music and speech activation.

Finally, no significant correlation was observed between participants' overall identification accuracy at screening (at criterion) and any of the decoding analyses nor did we observe any significant correlations between accuracy rates for just the instrument or speech

stimuli at screening and their respective within-category decoding analyses. Again, we also re-examined the univariate analysis and statistically controlled for these behavioral measures (as a covariates), but this also did not change the pattern of music and speech activation that we observed. We do note that initial overall screening task performance correlated negatively with maximum auditory localizer decoding ($r = –0.60$, $p = 0.009$), although this result is somewhat difficult to interpret as it suggests that better screening performance was associated with worse localizer decoding.

4. Discussion

Our results provide nuanced insight into whether separate or overlapping regions of cortex represent different individual sound sources that are important to human listeners. Individual sound sources from different categories (conspecific voices and musical instruments) were represented in largely separable regions of the superior temporal lobes, with just one voxel of overlap between categories. Decoding a stimulus' superordinate category ("music" or "speech") was more accurate and involved a larger area of superior temporal lobes. However, these representations were all couched within what appears to be a more general acoustic processing region that represents (at the very least) different aspects of the frequency spectrum.

Our tokens were selected so as to naturally vary in ways that facilitate natural speaker and instrument identification. Thus, our decoding results encompass acoustic feature differences within and between these categories and appear to reflect the brain's sensitivity to these dimensions. However, the minimal overlap between these categories suggests that early auditory areas (as defined by our localizer), which track these features, may be tuned to the acoustic properties specific to recognizing sound sources within each category. Thus, more work is

needed to determine whether the separate representations observed within these (localizer defined) auditory areas are a consequence of category-specific tuning to acoustic features among subsets of neurons within a sensory processing region (similar to Riesenhuber and Poggio, 2002), or if our results follow from acoustic feature differences inherent in our stimuli (similar to Giordano et al., 2014). Given the design of this study and stimulus set, we cannot rule out the latter possibility.

The spatial constraints of fMRI should be kept in mind when interpreting these overlap results. That is, "overlap" in a given region could be the result of separate but neighboring processes whose neurons overlap geographically within the space of voxel but do not interact functionally (e.g., Peretz et al., 2015). In this case there may be separate category specific and acoustic processes that occupy the same cortical region identified in our analyses, which are engaged to perform the different operations that we investigated. Perhaps these nuances can be explored by future work employing methods that are able to more finely decompose the neural responses within these areas (Norman-Haignere et al., 2015). Nevertheless, the pattern of overlap we observed is consistent with the idea that the distinct neural regions involved in processing speaker and timbre identity may be based on distinctions among more basic acoustic features processed by the brain.

Our speaker decoding results differ from some decoding analyses that identified larger portions of temporal cortex (Bonte et al., 2014; Formisano et al., 2008), but align more with other searchlight decoding results for conspecific voices (Hasan et al., 2016). This discrepancy suggests that speaker decoding may be based on a more distributed network of voxels than a given searchlight in this study might have had access to. Previous speaker decoding results have also used smaller voxel sizes (Formisano et al., 2008), suggesting that if the searchlight had

access to neural responses at a finer spatial resolution, speaker decoding might have been more accurate. Our speaker decoding results overlapped completely with our acoustic localizer and may have also been influenced by inherent differences in fundamental frequency, which is a strong cue for speaker identification (Creel and Bregman, 2011). Indeed, other studies of pitch processing implicate a similar region of auditory cortex (Penagos et al., 2004; Patterson et al., 2002).

The decoding of individual instruments from neural responses in fMRI has (to our knowledge) not been previously demonstrated, but the regions involved in instrument decoding were similar to previous results of studies that manipulated timbre features (Menon et al., 2002) or employed diverse stimulus sets involving changes in instrument timbre (Angulo-Perkins et al, 2014; Leaver and Rauschecker, 2010; Norman-Haignere et al., 2015). Taken together, these findings implicate the posterior and anterior temporal lobes, although our instrument decoding results largely overlapped with areas involved in decoding simple stimuli expressing low-level acoustic differences. This suggests that the neural representations of instruments that the classifier could discern may largely be based on acoustic differences, as others have suggested (Giordano et al., 2014). Interestingly, the decoding results obtained in our study echo findings from studies of timbre imagery (Halpern et al., 2004), which found that imagined timbres activate primary auditory cortex and posterior temporal lobes. Because imagery is not based on acoustic stimulation, this suggests that the representations of musical instruments we decoded in the posterior temporal lobes may relate to the same representations involved in timbre imagery. However, further work is needed to disentangle mental representations of instrument timbres and their perceptual or acoustic qualities.

We observed accurate category-level decoding that distinguished speech and musical sounds throughout the temporal lobes. This aligns with previous findings of both highly controlled exemplars of different categories (Staeren et al., 2009) and also with studies of diverse speech and non-speech sounds (Lee et al., 2015; Rogalsky et al., 2011). The between-category decoding results in the temporal lobes might have been partially related to the univariate differences in activation we observed between these categories, similar to previous studies identifying speech- (Overath et al., 2015) and music-specific regions of activation (Rogalsky et al., 2011). Taking our decoding and univariate results together with other findings of non-overlap between music and language (Fedorenko et al., 2012; Norman-Haignere et al., 2015; Rogalsky et al., 2011), it appears that speech and music sound sources are largely separable in their patterns of neural activation. Moreover, our tokens were simple, temporally limited examples of these sound sources (just 230 ms long), which presumably do not engage more complex structural or syntactic cognitive operations relevant to music or speech. Thus, these data suggest that the processing of music and speech are separable even in very early stages of cortical processing, when the auditory system is identifying sound sources and processing acoustic cues, prior to more cognitively complex linguistic or music functions.

It is notable that we were able to achieve reliable decoding of the auditory localizer stimuli despite having just three runs for stimulus presentation, which constituted a more limited set of training data than the searchlights in the main study. This could be due to both the simple nature of the stimuli and our use of ISSS (Schwarzbauer et al., 2006) which likely improved the temporal signal-to-noise ratio in our analyses (Murphy et al., 2007; Peelle, 2014). We note however, that we did not attempt to specifically delineate different tonotopic axes, which likely would require more data (cf. Humpheries et al., 2010). Rather, our aim was simply to identify

regions which might be involved in simple auditory processing operations. Nonetheless, our results suggest that ISSS may be an effective and efficient way to probe such low-level auditory processes in future work.

Finally, we examined the relationship between our fMRI results and musical training as well as behavioral identification accuracy for the stimuli as assessed at screening. Musical training was associated with a benefit to initial screening performance for instruments (only). However, four participants repeated the screening task and this extra exposure appears to have overcome any potential differences in familiarity associated with musical training. Indeed, all participants reached the high accuracy criterion we set in order to proceed to the scanning session (> 88 % accuracy overall) and, at criterion, there was no correlation between musical training and task performance. This is therefore unlikely to have had a substantial influence on our fMRI results, since musical training correlated not with musical instrument decoding accuracy (where experience might be hypothesized to strengthen neural representations), but with speaker decoding accuracy and did not influence the music and speech findings in the univariate analysis. When taken with the similarities between the speaker decoding cluster and classic pitch processing regions, this suggests that musical training may be related to the degree of reliance on particular acoustic cues (e.g., pitch) for learning and representing individual speakers.

However, the influence of the individual musical training factors and behavioral correlates in these results are tentative and should be more thoroughly explored by future work for a number of reasons. First, our participant sample was small relative to studies of musical training effects and likely does not fully represent the spectrum of musical training necessary to make strong conclusions about the relationship between musical experience and these auditory

processes. Second, we required a high level of behavioral performance in order to proceed to the fMRI session. This was to in order to ensure that participants could indeed perceptually discriminate the acoustic qualities of the stimuli, which would in turn indicate some level of separability in participants' auditory responses that the classification analysis might detect. However, this may have imposed a range restriction on the behavioral screening performance measures. Thus, follow up work involving a greater number of participants with a wider range of musical ability and a larger, more difficult stimulus set would allow for a more nuanced and in-depth examination of the topics explored in the correlation analyses.

It is also worth noting two more general limitations of our design and how they are unlikely to influence our findings. First, we used a relatively short trial length of 8 seconds to maximize the number of trials for training and testing the classifier. Such event-related fMRI designs create the potential for BOLD activation from a previous trial to extend into a subsequent trial. However, the order of stimuli in each list was randomized and no individual sound source was played on any two consecutive trials so this could not have favored any particular sound or classification in our results. The second potential source of noise is the T1-signal decay artifact from the ISSS sequence that we observed on the first volume after the silent TR. A visualization of this issue is shown in Supplementary Figure 1 where it can be seen that TRs 2 and 3 were unaffected. Given that this artifact was easily managed by removing the first volume in each trial, we view this as an acceptable shortcoming given that ISSS allowed for sounds to be presented in quiet, while increasing the temporal signal-to-noise ratio of the data. Moreover, these issues do not confound our results because they were either present on all trials (T1 decay) or were randomly distributed across them (trial order). Finally, to further guard

against noise from spuriously influencing our results, classification was assessed using empirically derived chance distributions based on null permuted data sets (Stelzer et al., 2013).

The present study set out to probe whether the brain represents sound sources within and between different categories using overlapping or separable neural substrates. We found that the voices of conspecifics and the sounds of different musical instruments were successfully decoded from other sounds from the same category in separable regions of superior temporal cortex in and around primary auditory cortex. Decoding sound category membership (speaker or instrument) was more accurate and involved a more extensive portion of the temporal lobes. Importantly, the results of an auditory localizer that used simple, acoustically separable stimuli indicated that most of the decoding we observed occurred in regions that are also involved in representing simple acoustic differences. The results we report here help fill a critical gap in our understanding of how human listeners so efficiently identify the sounds in their auditory environment, which in turn informs later, downstream processes, such as auditory scene analysis, as well as speech and music perception. This investigation and future studies can guide the development of automated sound source identification algorithms and assistive therapies.

5. Acknowledgments

This work was supported by a seed grant from the Center for the Advanced Study of Language at University of Maryland. The authors would like to thank Ed Smith for his help optimizing the audio presentation and playback system in the scanner.

6. References

Akama, H., Murphy, B., Na, L., Shimizu, Y., and Poesio, M. (2012). Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. *Frontiers in Neuroinformatics*, *6*, 24.

Allen, E. J., Burton, P. C., Olman, C. A., and Oxenham, A. J. (2017). Representations of pitch and timbre variation in human auditory cortex. *Journal of Neuroscience, 37*, 1284–1293.

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*, 3677–3689.

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., and Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes: Differences between musicians and non-musicians. *Cortex*, *59*, 126–137.

Belin, P., and Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport*, *14*, 2105–2109.

Bizley, J. K., and Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*, 693–707.

Bonte, M., Hausfeld, L., Scharke, W., Valente, G., and Formisano, E. (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *Journal of Neuroscience*, *34*, 4548–4557.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge: MIT Press.

Chandrasekaran, B., Chan, A. H., and Wong, P. C. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*, 2690–2700.

Chartrand, J. P., and Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters, 405*, 164–167.

Chen, G., Saad, Z., Nath, A., Beauchamp, M., and Cox, R. W. (2012). Fmri group analysis combining effect estimates and their variances. *Neuroimage, 60*, 747–765.

Corrigall, K. A., Schellenberg, E. G., and Misura, N. M. (2013). Music training, cognition, and personality. *Frontiers in Psychology*, *4*, 222.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*, 162–173.

Cox, R. W., Chen, G., Glen, D. R., Reynolds, R. C., and Taylor, P. A. (2017). FMRI clustering in AFNI: False-positive rates redux. *Brain Connectivity*, *7*, 152–171.

Creel, S. C., and Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass, 5*, 190–204.

De Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*, 1917–1930.

Fedorenko, E., McDermott, J. H., Norman-Haignere, S., and Kanwisher, N. (2012). Sensitivity to musical structure in the human brain. *Journal of Neurophysiology*, *108*, 3289–3300.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*, 970–973.

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., and Belin, P. (2012). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, *23*, 2025–2037.

Giordano, B. L., Pernet, C., Charest, I., Belizaire, G., Zatorre, R. J., and Belin, P. (2014). Automatic domain-general processing of sound source identity in the left posterior middle frontal gyrus. *Cortex*, *58*, 170–185.

Halpern, A. R., Zatorre, R. J., Bouffard, M., and Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*, 1281–1292.

Hasan, B. A. S., Valdes-Sosa, M., Gross, J., and Belin, P. (2016). "Hearing faces and seeing voices": Amodal coding of person identity in the human brain. *Scientific Reports*, *6*, 37494.

Haynes, J. D., and Rees, G. (2006). Neuroimaging: Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience, 7*, 523–534.

Hjortkjær, J., Kassuba, T., Madsen, K. H., Skov, M., and Siebner, H. R. (2017). Task-Modulated Cortical Representations of Natural Sound Source Categories. *Cerebral Cortex*, *28*, 295–306.

Humphries, C., Liebenthal, E., and Binder, J. R. (2010). Tonotopic organization of human auditory cortex. *NeuroImage*, *50*, 1202–1211.

Jimura, K., and Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*, 544–552.

Khalighinejad, B., da Silva, G. C., and Mesgarani, N. (2017). Dynamic encoding of acoustic features in neural responses to continuous speech. *Journal of Neuroscience*, *37*, 2176–2185.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences, 103*, 3863–3868.

Leaver, A. M., and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: Effects of acoustic features and auditory object category. *Journal of Neuroscience*, *30*, 7604–7612.

Lee, Y. S., Peelle, J. E., Kraemer, D., Lloyd, S., and Granger, R. (2015). Multivariate sensitivity to voice during auditory categorization. *Journal of Neurophysiology*, *114*, 1819–1826.

McAdams, S., and Giordano, B. L. (2009). The perception of musical timbre. In S. Hallam, I. Cross, and M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 72–80). Oxford, UK: Oxford University Press.

Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., ... and McAdams, S. (2002). Neural correlates of timbre change in harmonic sounds. *NeuroImage*, *17*, 1742–1754.

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS One, 9*, e89642.

Murphy, K., Bodurka, J., and Bandettini, P. A. (2007). How long to scan? The relationship between fMRI temporal signal to noise ratio and necessary scan duration. *Neuroimage*, *34*, 565–574.

Norman-Haignere, S., Kanwisher, N. G., and McDermott, J. H. (2015). Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, *88*, 1281–1296.

Ogg, M., and Slevc, L. R. (*in press*). Neural mechanisms of music and language. In G. Zubicaray, and N.Schiller (Eds.), *Oxford Handbook of Neurolinguistics*.

Ogg, M., Slevc, L. R., and Idsardi, W. J. (2017). The time course of sound category identification: Insights from acoustic features. *The Journal of the Acoustical Society of America, 142*, 3459–3473.

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I. H., Saberi, K., ... and Hickok, G. (2010). Hierarchical organization of human auditory cortex: Evidence from acoustic invariance in the response to intelligible speech. *Cerebral Cortex*, *20*, 2486–2495.

Oosterhof, N. N., Connolly, A. C., and Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*, 27.

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature Neuroscience*, *18*, 903– 911.

Owren, M. J., and Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America*, *119*, 1727–1739.

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., and Griffiths, T. D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron, 36*, 767–776.

Peelle, J. E. (2014). Methodological challenges and solutions in auditory functional magnetic resonance imaging. *Frontiers in Neuroscience*, *8*, 253.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13.

Penagos, H., Melcher, J. R., and Oxenham, A. J. (2004). A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. *Journal of Neuroscience, 24*, 6810–6815.

Peretz, I., Vuvan, D., Lagrois, M. É., & Armony, J. L. (2015). Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *370*, 20140090.

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K., and Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences, 19*, 783–796.

Riesenhuber, M., and Poggio, T. (2002). Neural mechanisms of object recognition. *Current Opinion in Neurobiology, 12*, 162–168.

Rogalsky, C., Rong, F., Saberi, K., and Hickok, G. (2011). Functional anatomy of language and music perception: Temporal and structural factors investigated using functional magnetic resonance imaging. *Journal of Neuroscience*, *31*, 3843–3852.

Schwarzbauer, C., Davis, M. H., Rodd, J. M., and Johnsrude, I. (2006). Interleaved silent steady state (ISSS) imaging: A new sparse imaging method applied to auditory fMRI. *NeuroImage*, *29*, 774–782.

Smith, S. M., and Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*, 83–98.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., and Formisano, E. (2009). Sound categories are represented as distributed patterns in the human auditory cortex. *Current Biology*, *19*, 498–502.

Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage*, *65*, 69–82.

Tong, F., and Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology*, *63*, 483–509.

Von Kriegstein, K., Eger, E., Kleinschmidt, A., and Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*, 48–55.

Von Kriegstein, K., Kleinschmidt, A., Sterzer, P., and Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*, 367–376.

Warren, J. D., Jennings, A. R., and Griffiths, T. D. (2005). Analysis of the spectral envelope of sounds by the human brain. *NeuroImage*, *24*, 1052–1057.

Zatorre, R. J., Bouffard, M., and Belin, P. (2004). Sensitivity to auditory object features in human temporal neocortex. *Journal of Neuroscience*, *24*, 3637–3642.

Zendel, B. R., and Alain, C. (2012). Musicians experience less age-related decline in central auditory processing. *Psychology and Aging*, *27*, 410–417.