The MIT Press

RESEARCH ARTICLE

Right Posterior Temporal Cortex Supports Integration of Phonetic and Talker Information

Sahil Luthra[1] [iD], **James S. Magnuson**[1,2,3], and **Emily B. Myers**[1,4]

[1]Department of Psychological Sciences, University of Connecticut, Storrs, CT, USA
[2]Basque Center on Cognition Brain and Language (BCBL), Donostia-San Sebastián, Spain
[3]Ikerbasque, Basque Foundation for Science, Bilbao, Spain
[4]Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, CT, USA

ABSTRACT

Though the right hemisphere has been implicated in talker processing, it is thought to play a minimal role in phonetic processing, at least relative to the left hemisphere. Recent evidence suggests that the right posterior temporal cortex may support learning of phonetic variation associated with a specific talker. In the current study, listeners heard a male talker and a female talker, one of whom produced an ambiguous fricative in /s/-biased lexical contexts (e.g., *epi?ode*) and one who produced it in /ʃ/-biased contexts (e.g., *friend?ip*). Listeners in a behavioral experiment (Experiment 1) showed evidence of lexically guided perceptual learning, categorizing ambiguous fricatives in line with their previous experience. Listeners in an fMRI experiment (Experiment 2) showed differential phonetic categorization as a function of talker, allowing for an investigation of the neural basis of talker-specific phonetic processing, though they did not exhibit perceptual learning (likely due to characteristics of our in-scanner headphones). Searchlight analyses revealed that the patterns of activation in the right superior temporal sulcus (STS) contained information about who was talking and what phoneme they produced. We take this as evidence that talker information and phonetic information are integrated in the right STS. Functional connectivity analyses suggested that the process of conditioning phonetic identity on talker information depends on the coordinated activity of a left-lateralized phonetic processing system and a right-lateralized talker processing system. Overall, these results clarify the mechanisms through which the right hemisphere supports talker-specific phonetic processing.

INTRODUCTION

Speech scientists have long appreciated that individual talkers can differ considerably in how they produce their speech sounds, such that the acoustics that signal an /ɛ/ (the vowel in "bed") for one talker might correspond to an /æ/ (as in "bad") for another (Hillenbrand et al., 1995; Joos, 1948; Peterson & Barney, 1952). Listeners are sensitive to the acoustic-phonetic variability across talkers (Allen & Miller, 2004; Newman et al., 2001; Theodore & Miller, 2010) but nevertheless are able to accurately recognize words spoken by a wide range of talkers (e.g., Liberman et al., 1967). Thus, there must be some mechanism by which listeners accommodate talker variability—that is, some way in which they can condition their interpretation of a speech sound on their knowledge of who is talking. Bayesian accounts posit that listeners accommodate talker variability by maintaining distinct sets of beliefs—that is, a

Generative model:
Beliefs (likely implicit) about how certain talkers produce their speech sounds; sensitivity to acoustic patterns that correlate with talker/phonetic identity.

distinct generative model—for how a given talker produces speech sounds; a model may describe an individual talker or a group of talkers that share socio-indexical traits such as age, gender, sexual orientation, and/or national origin (Kleinschmidt, 2019).

High-level information, such as word-level knowledge, plays a critical role in helping listeners adapt to the idiosyncratic ways that different talkers produce speech (Maye et al., 2008; Norris et al., 2003). In a landmark study, Norris et al. (2003) exposed Dutch listeners to a talker who often produced a speech sound that was ambiguous between /s/ and /f/ (denoted /ʔ/). Crucially, one group of listeners only heard the ambiguous phoneme in contexts where lexical information biased their interpretation of the phoneme toward /s/ (contexts like [ra:dɛiʔ], where [ra:dɛis] (*radijs*) is the Dutch word for radish, but [ra:dɛif] (**radijf*) is a nonword), and the other group only heard the ambiguous phoneme in /f/-biased contexts. During an initial exposure phase, listeners were asked to indicate whether these items were real words or nonwords, and they consistently endorsed items with ambiguous phonemes as being real words, suggesting that they used lexical knowledge to guide their immediate interpretation of the ambiguous phoneme (an effect previously shown by Ganong, 1980). After the initial exposure phase, Norris et al. had participants complete a phonetic categorization task with stimuli from an /f/-/s/ nonword–nonword continuum produced by the same talker. They found that participants who had previously heard ambiguous fricatives in /s/-biased contexts were likely to identify similar ambiguous fricatives as /s/, and those who had heard ambiguous fricatives in /f/-biased contexts were more likely to interpret ambiguous tokens as /f/. That is, contexts encountered during exposure allowed listeners to update their beliefs for how that talker produced speech sounds. Work on lexically guided perceptual learning has shown that listeners can simultaneously track the phonetic idiosyncrasies of two different talkers (e.g., if one produces an ambiguous fricative in place of /s/ while another produces an ambiguous fricative in place of /ʃ/ ["sh"]; Kraljic & Samuel, 2005, 2007; Luthra et al., 2021), consistent with the proposal that listeners can maintain separate generative models for different sets of talkers.

Lexically guided perceptual learning:
Paradigm wherein listeners use vocabulary knowledge to update their generative model of how some talker produces their speech sounds.

Some insight into the neural mechanisms through which listeners contact generative models to guide speech perception comes from a lexically guided perceptual learning study conducted by Myers and Mesite (2014). The authors found that during phonetic categorization, two regions exhibited differential responses to the ambiguous tokens as a function of previous exposure—the right middle frontal gyrus (MFG) and the right middle temporal gyrus (MTG). The involvement of the right hemisphere may be surprising, given that phonetic processing is primarily supported by regions in the left hemisphere (Rauschecker & Scott, 2009; Turkeltaub & Branch Coslett, 2010). However, prominent models of speech perception, such as the dual stream model (Hickok & Poeppel, 2000, 2004, 2007) point out that at low levels of the language hierarchy (phonology), information may be fairly bilaterally represented, and right temporal regions are often recruited in functional neuroimaging studies of speech perception (Belin et al., 1999; Blumstein et al., 2005; Davis et al., 2011; Giraud et al., 2004; Zatorre et al., 1996). Furthermore, vocal identity processing relies principally on the contributions of right temporal cortex (Belin et al., 2000; Belin & Zatorre, 2003; Schall et al., 2014; von Kriegstein et al., 2003; von Kriegstein & Giraud, 2004) and sometimes places demands on right frontal cortex as well (Andics et al., 2013; Stevens, 2004). Myers and Mesite therefore suggested that the involvement of the right hemisphere in their study might reflect the fact that during the phonetic categorization runs, listeners had to access their beliefs about the idiosyncratic way that this particular talker produced speech sounds.

In a follow-up study, Luthra et al. (2020) performed multivoxel pattern analyses (MVPA) on the data from the lexically guided perceptual learning study by Myers and Mesite (2014) described above. In particular, Luthra et al. trained a machine learning algorithm on the

correspondences between phonetic identity and patterns of functional activation for unambiguous stimuli (i.e., clear productions of /asi/ and /aʃi/ that participants heard during the phonetic categorization runs). When the classifier was tested on the patterns of functional activation for the ambiguous stimuli (i.e., those in the middle of the /asi/-/aʃi/ continuum), with trials labeled based on whether the listener had reported /s/ or /ʃ/ on that particular trial, the classifier achieved above-chance accuracy. That is, the pattern of activation for an ambiguous trial resembled the canonical pattern for whichever endpoint (/s/ or /ʃ/) the subject reported having heard. Their initial analysis considered a broad set of regions involved in language processing, but exploratory region-of-interest (ROI) analyses found that above-chance accuracy was still observed when the classifier only received information about the activity of left parietal cortex or information about right temporal cortex. These findings provide further evidence for a right hemisphere role—right temporal cortex in particular—in representing a listener's perceptual interpretation of speech from a talker with systematically atypical productions. Notably, however, listeners in the Myers and Mesite study only heard speech from one talker, so it is unclear whether the involvement of the right temporal cortex is a consequence of listeners contacting a *talker-specific* generative model.

The proposal that phonetic information and talker detail might be integrated in the right temporal cortex is consistent with a growing body of evidence that the right temporal cortex simultaneously represents phonetic detail and talker information (Evans & Davis, 2015; Formisano et al., 2008; von Kriegstein et al., 2010). In a seminal study, Formisano et al. (2008) presented listeners with three different vowels produced by three different talkers while measuring brain responses using fMRI. Using a machine learning technique, they found that the most informative voxels for phonetic classification spanned bilateral temporal areas including the superior temporal gyrus (STG), superior temporal sulcus (STS), and MTG. In contrast, the most informative voxels for talker identity were right-lateralized, located primarily in the right STS, though a small number of voxels in left STS also contributed meaningfully. Crucially, there was a small subset of voxels in which activation patterns were useful in both the classification of phonetic identity *and* of talker identity, with these voxels primarily located in the right STS, though some degree of overlap in the left STS was also observed. The authors emphasized the finding that most voxels that were informative for phonetic classification were not informative for talker classification (and vice versa), suggesting that phonetic information and talker information are segregated relatively early in processing.

Additional evidence for this view comes from a functional MRI study by Evans and Davis (2015). To investigate how different dimensions of the speech signal are encoded across the brain, the authors used representational similarity analyses (Kriegeskorte et al., 2008) to examine the similarity of functional activation patterns as syllable identity, talker identity, and degree of acoustic degradation were parametrically manipulated. The authors found that the left and right temporal cortex—specifically, bilateral clusters encompassing the STG, STS, and MTG—responded similarly to trials where syllable identity and talker identity were held constant, regardless of the kind of acoustic degradation in the speech signal. Thus, the bilateral temporal cortex may specifically support the integration of phonetic identity and talker identity, as changes to either one of these dimensions can affect the pattern of functional activity.

However, it is notable that the stimuli used in these studies were relatively unambiguous, where there might be little pressure to link phonetic tokens to a talker identity. In order to clarify the mechanisms through which phonetic details and talker information are integrated, it may be important to consider the case of phonetically ambiguous stimuli, where listeners must appeal to talker information in order to resolve phonetic identity. Furthermore, the finding that a set of right STS voxels was highly discriminative for both phonetic discrimination and

talker discrimination provides a clue that the right STS may be an important interface between a left-lateralized phonetic processing system and a right-lateralized talker processing system.

In concert, the findings from these studies suggest that some set of voxels in the temporal cortex may serve as an interface between the left-lateralized phonetic processing system and the right-lateralized talker processing system. However, there are a number of important details that need to be clarified. While the results of Formisano et al. (2008) implicate the STS in particular, the results of Evans and Davis (2015) suggest that the STG and MTG may also contribute meaningfully. As such, an important next step is to more precisely characterize the contributions of the STG, STS, and MTG in integrating phonetic detail with talker information (Figure 1).

In addition to the possibility that phonetic information and talker information are simultaneously represented in some part of the bilateral temporal lobes (i.e., that there is some degree of *overlap* between the phonetic processing system and the talker processing system), it is also possible that talker-specific phonetic processing is partly achieved through functional connections between the left and right temporal cortex (i.e., that there is some degree of *interaction* between the two neural systems). Evidence that the integration of phonetic information and talker information is supported by interhemispheric interactions comes from a study by von Kriegstein et al. (2010). In that study, participants listened to a series of syllables produced by talkers who differed in vocal tract length. Vocal tract length is a parameter that affects the formants (i.e., the frequency bands where acoustic energy is most concentrated) in the speech signal, and as such, it can systematically influence the acoustic information associated with different phonemes (Johnson, 2008; Joos, 1948) as well as different talkers (Fitch & Giedd, 1999). Strikingly, von Kriegstein et al. found that the response of the right posterior STG/STS was influenced by vocal tract length, but only when listeners were engaged in a speech perception task (a one-back syllable monitoring task) and not in a control task (one-back loudness monitoring or one-back talker monitoring). The authors suggested that the right posterior STG/STS activity might reflect the estimation of vocal tract length for the purpose of talker normalization; that is, the recruitment of the right STG/STS might underlie the process by which listeners leverage their knowledge of the talker's formant structure to make adjustments to the mapping from acoustics to phonetic categories. This finding implicates the right STG/STS in the integration of phonetic detail and talker information. However, the authors also found a cluster of voxels in the left STG that was sensitive to vocal tract length, and critically, the activity of this left STG cluster was strongly correlated with the activity of an analogous STG cluster in the right hemisphere during the speech recognition task; that is, the two regions were functionally connected. Notably, the functional connectivity between the left and right STG was

Functional connectivity analysis: Analysis examining how well activity in a seed region predicts activity of other regions; assesses interactions between multiple brain regions.
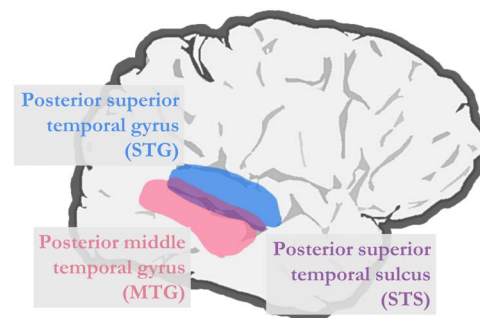
Figure 1. Right temporal brain regions that may support the integration of phonetic detail with talker information. Previous studies have implicated the right posterior STG and MTG, as well as the right STS, which lies between the STG and MTG, in conditioning phonetic identity on talker information. However, the precise contribution of each of the different brain regions remains underspecified.

significantly weaker when listeners were performing the control tasks. Taken together, the data from von Kriegstein et al. suggest that in order to leverage talker information for the purposes of speech perception, listeners rely on both the left and right temporal cortex. That is, the integration of talker detail and phonetic information may be partly supported by interactions between right hemisphere regions associated with talker processing and left hemisphere regions associated with phonetic processing.

The goal of the current fMRI study was to provide a more precise characterization of the neural mechanisms by which listeners contact talker-specific generative models. We specifically sought to examine the (implicit) beliefs that form over the course of lexically guided perceptual learning; in doing so, we hoped to better characterize the neural systems that listeners use when updating their beliefs about how different talkers produce their speech sounds. We considered two possible neural mechanisms. First, we considered the possibility that phonetic detail and talker information are simultaneously encoded in a single brain region, representing an overlap between the neural system for phonetic processing and the neural system for talker processing. Here, we focused specifically on the right temporal cortex, which previous studies have suggested may support the integration of talker information and phonetic detail (Evans & Davis, 2015; Formisano et al., 2008). Second, we considered the degree to which talker-specific phonetic processing is achieved by the coordinated activity of the left and right temporal cortex—that is, by functional interactions between the two neural systems. Importantly, these two mechanisms are not mutually exclusive, and the process of leveraging talker information for phonetic processing may depend on both.

The approach of our study was as follows. Over the course of several exposure runs, listeners were exposed to two talkers, one of whom produced an ambiguous fricative (a blend between /s/ and /ʃ/) in /s/-biased contexts like *epi?ode* and the other of whom produced the same ambiguous fricative in /ʃ/-biased contexts like *friend?ip*. Thus, in these runs, lexical information would encourage listeners to form talker-specific beliefs of how acoustic-phonetic information maps onto the /s/ and /ʃ/ sounds. In test runs, listeners heard stimuli from a /s/-/ʃ/ continuum and performed a phonetic categorization task. Critically, to interpret ambiguous phonemes, listeners had to condition phonetic identity on talker information (talker identity or acoustic patterns that covary with talker). Neuroimaging data were collected using fMRI. To assess whether any region integrates phonetic information and talker information, we conducted a series of searchlight analyses to identify voxels where the local pattern of activation contained information about the phonetic identity of the test stimulus as well as about the talker who produced it. To measure the extent to which phonetic information and talker information are integrated through interactions between distinct brain areas, we conducted a functional connectivity analysis. The stimuli and analysis code for this work are publicly available at https://osf.io/6j7nr/.

We first performed a behavioral experiment (Experiment 1) aimed at confirming that listeners could simultaneously maintain distinct sets of beliefs for our two talkers. Once we established that this was the case, we conducted an fMRI experiment (Experiment 2) aimed at clarifying the contributions of different neural regions when listeners must condition phonetic identity on talker information.

Searchlight analysis:
Form of multivoxel pattern analysis wherein a roving pattern classifier only considers the local pattern of activation.

EXPERIMENT 1

Previous lexically guided perceptual learning studies have established that listeners are able to maintain distinct sets of beliefs for two different talkers (e.g., one who produces an ambiguous /s/-/ʃ/ blend in /s/-biased contexts like *epi?ode* and one who produces an ambiguous /s/-/ʃ/

blend in /ʃ/-biased contexts such as *friend?ip*; Kraljic & Samuel, 2007; Luthra et al., 2021). In Experiment 1, we sought to verify that we could induce talker-specific perceptual learning, adapting the approach of Luthra et al. (2021) to meet the design constraints of our fMRI experiment (Experiment 2). Thus, Experiment 1 provides a baseline characterization of the learning effects that we would expect to observe in Experiment 2.

Materials and Methods

Stimuli

Stimuli were taken from Luthra et al. (2021), to which the reader is directed for additional information on stimulus construction. In brief, the stimuli consisted of a set of items for exposure runs and a separate set of items for test runs. The exposure items consisted of 32 words, 16 containing a word-medial /s/ and 16 containing a word-medial /ʃ/. These two sets of items were matched in word frequency, number of syllables prior to the fricative, and number of total syllables. A female native speaker of American English produced lexically consistent (e.g., *episode*) and lexically inconsistent (e.g., *epishode*) versions of each item. Word-nonword continua were made for each item using STRAIGHT (Kawahara et al., 2008), and for each item, the authors selected a continuum step with an unambiguous fricative (e.g., *episode*, *friendship*) and a step with an ambiguous fricative (*epi?ode*, *friend?ip*). Male versions of these stimuli were created by applying the "Change Gender" function in Praat (Boersma & Weenik, 2017), which uses a Pitch Synchronous Overlap and Add (PSOLA) algorithm to shift the pitch of a stimulus and adjust the formant ratio; the pitch change captures the general phenomena that male voices tend to have more massive vocal folds and thus lower pitch (e.g., Slavit, 1999), whereas the formant ratio is related to vocal tract length, which is in turn correlated with body size (e.g., Fitch & Giedd, 1999). For the test phase, the female talker was recorded saying the words *sign* and *shine*; a 7-step continuum was created in STRAIGHT, and male versions of the stimuli were created using the "Change Gender" function as described above.

For the present experiment, we opted to use only four steps from each test continuum, following the approach of Myers and Mesite (2014). Pilot testing with these stimuli indicated that steps 1 and 6 were relatively unambiguous, with participants interpreting the initial segment as /ʃ/ 2% and 98% of the time, respectively. Steps 3 (34% /ʃ/ response) and 4 (76% /ʃ/) were selected as the ambiguous step, and both steps were associated with sizable shifts in perception based on previous exposure.

Participants

Twenty-six individuals were recruited using the online participant recruitment platform Prolific (Prolific Academic, 2022). All participants were English-speaking monolinguals residing in the United States between the ages of 18 and 34. All participants had normal or corrected-to-normal vision, no hearing loss, and no language-related disorders. Individuals who had participated in previous studies using these stimuli were not eligible for the current experiment, and only participants using a desktop computer were able to complete the experiment. All participants provided informed consent, and each participant was paid $7.50 for their time. All procedures were approved by the University of Connecticut Institutional Review Board.

Data from three participants were excluded due to technical errors. We decided a priori to exclude the data of any participants who failed the headphone screening test (described below) twice, resulting in the exclusion of one more participant. Luthra et al. (2021) also excluded participants who failed to respond to 10% or more of the trials in either the exposure or test runs as well as participants who showed less than 70% accuracy in classification of the

unambiguous endpoints during the phonetic categorization task. Applying this criterion led to the exclusion of two additional participants. After these exclusionary criteria were applied, data from 20 participants (16 female, 4 male; mean age: 26, age range: 20–31) were included in the analyses for Experiment 1.

Procedure

After providing informed consent, participants completed a screening test designed to verify that they were using headphones (Woods et al., 2017). In this test, a sequence of auditory tones is presented binaurally, and participants are asked to indicate which tone is quietest. Critically, one tone in the sequence is presented in anti-phase across the stereo channels. When presented over loudspeakers, waveforms from the left and right channels cancel centrally, resulting in an attenuated signal; thus, participants completing this task over loudspeakers generally tend to select this tone. Over headphones, however, the anti-phase tone is not heard as the quietest. Thus, participants are expected to perform differently depending on whether they are wearing headphones, and this can be used to screen participants thought to be using their computer's loudspeakers. After completing the headphone screening test, participants answered a short set of demographics questions and then moved on to the main experimental task.

As schematized in Figure 2, the experimental task alternated between exposure runs and test runs. During each exposure run, listeners heard speech from one of the two talkers, and heard ambiguous fricatives from that talker in lexical contexts that led to the interpretation as either /s/ or /ʃ/, with the particular bias of the talker (i.e., whether the male or female talker was /s/- or /ʃ/-biased) counterbalanced across participants. For instance, if the talker was /s/-biased, the listener would hear 16 ambiguous fricatives in contexts where lexical information biased them to interpret the segment as /s/ (e.g., *epi?ode*) and 16 clear productions of the contrastive category (e.g., *refreshing* with a clear /ʃ/). In test runs, listeners performed a phonetic categorization task with stimuli from a *sign–shine* continuum. In each test run, stimuli were produced by the same talker who had produced the stimuli during the previous exposure run, and listeners heard 12 instances of each of the four steps (two unambiguous, two ambiguous). After performing an exposure run and a test run with one talker, listeners would perform an exposure run and a test run with the other talker. This set of four runs was repeated four times over the course of the experiment, for a total of 16 runs. The order of the talkers (male first or female first) was also counterbalanced across participants.

During exposure trials, listeners simply had to indicate whether the talker they were hearing was a male talker or a female talker. Critically, because talker was held constant in each run, participants made the same response for the entirety of each exposure run. Our goal in
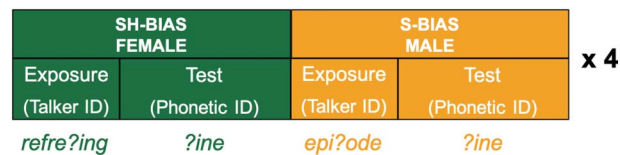
SH-BIAS FEMALE		S-BIAS MALE		
Exposure	Test	Exposure	Test	**x 4**
(Talker ID)	(Phonetic ID)	(Talker ID)	(Phonetic ID)	
refre?ing	*?ine*	*epi?ode*	*?ine*	

Figure 2. Overview of the design for Experiments 1 and 2. Listeners alternated between exposure runs (during which they performed a talker identification task) and test runs (during which they performed a phonetic categorization task). During an exposure run, listeners would hear one of the two talkers (e.g., the female talker) producing ambiguous fricatives (noted /?/) in lexically biased contexts (e.g., /ʃ/-biased contexts like *friend?ip*). During a subsequent test run, listeners heard stimuli from a *sign–shine* continuum, with all items produced by the same talker as in the previous exposure run. Listeners completed a total of 16 experimental runs. We counterbalanced which talker was associated with which biasing condition as well as which talker listeners heard first.

adopting such a simple task was to consistently remind listeners of the association between the phonetic information and the talker, potentially encouraging the formation of talker-specific generative models. Notably, lexically guided perceptual learning has been observed with a variety of exposure tasks, several of which have required relatively shallow processing (Drouin & Theodore, 2018; Eisner & McQueen, 2006; Luthra et al., 2021; Maye et al., 2008; McQueen et al., 2006; White & Aslin, 2011). Furthermore, previous work has suggested that attending to talker identity encourages listeners to encode talker-specific phonetic detail (Goldinger, 1996; Theodore et al., 2015). Finally, previous research has demonstrated that the degree to which right temporal regions respond to speech is modulated by the extent to which listeners attend to talker identity (Schall et al., 2014; von Kriegstein et al., 2003), motivating the use of a talker identification task for the current study.

Because the overall goal of the present study was to investigate the neural mechanisms through which listeners contact talker-specific generative models, we are particularly interested in characterizing neural activity during the test phase, during which listeners must leverage talker information to perform phonetic classification. As such, we wanted to have a relatively large number of test trials. However, because listeners continually update their beliefs of how a particular talker produces their speech sounds (Saltzman & Myers, 2021; Tzeng et al., 2021), we were concerned that hearing unambiguous fricatives during a protracted test phase might encourage listeners to abandon the beliefs formed during the exposure phase. It was for this reason that we opted to have listeners in this study alternate between exposure runs and test runs, following the approach of Myers and Mesite (2014). In this way, a listener's beliefs about how the talker produced their speech sounds would be re-established prior to each test phase. Note that in theory, listeners could show what might appear to be talker-specific learning if they just adjusted their category boundary based on the most recent exposure condition, without actually forming talker-specific beliefs; however, results from previous lexically guided perceptual learning studies with multiple voices suggest that listeners do establish and maintain talker-specific beliefs for multiple talkers, rather than simply being guided by their most recent exposure block (Kraljic & Samuel, 2007; Luthra et al., 2021).

Across participants, we counterbalanced which talker listeners heard first (the female or the male) as well as the bias (/s/-bias or /ʃ/-bias) assigned to each talker. Within each run, trial order was randomized for each participant, and response mappings were counterbalanced across participants. Participants had 4 s to respond on each trial, and each trial was followed by a 1 s intertrial interval. Experiment 1 was programmed using the Gorilla online experiment builder (Anwyl-Irvine et al., 2020).

Analyses

Trial-level phonetic categorization data were analyzed with a mixed-effects analysis implemented in R (Version 4.2.1; R Core Team, 2022). We specifically used the "mixed" function in the afex package (Singmann et al., 2018); this function implements mixed-effects models using the "glmer" function of the lme4 package (Bates et al., 2015) and assesses the significance of each fixed effect using likelihood ratio tests. We specified a logit link for our analyses, as appropriate for dichotomous data.

Our model attempted to predict the likelihood of a subject making a "shine" response and considered fixed factors of Step (scaled), Bias (sum-coded [1, −1], sh-bias, s-bias), and Talker (sum-coded [1, −1], female, male). The model also included random intercepts for each subject, random by-subject slopes for Step, Bias, and Talker, and random by-subject interactions

between Step and Bias as well as between Step and Talker. (Because Bias and Talker were consistently paired for a given individual—that is, a given listener either heard an /s/-biased female talker or an /ʃ/-biased female talker, but never both—we did not include random by-subject interactions between Bias and Talker.) Thus, the full model syntax was specified as:

```
SH_resp ~ step * bias * talker + (step:bias + step:talker + step + bias +
talker | subject)
```

This represents both the maximal model (Barr et al., 2013) and the most parsimonious, as a backward-stepping procedure indicated that a simpler random effect structure yielded a significantly worse model fit (Matuschek et al., 2017).

Results

Phonetic categorization data are plotted in Figure 3. Visually, it appears that listeners' responses were strongly affected by whether the talker had previously produced the ambiguous sound in /ʃ/-biased or /s/-biased contexts (left panel) but not strongly affected by the gender (male or female) of the talker (right panel).

Results of the statistical analysis are given in Table 1. We observed an expected main effect of Step, as participants made more "shine" responses when presented with steps closer to the "sh" end of the continuum. There was a significant main effect of Bias, with participants making more "sh" responses if they had previously heard the ambiguous sound in contexts where /ʃ/ was the lexically consistent phoneme; that is, we observed lexically guided perceptual learning. There was also a marginal Step × Bias interaction ($p = 0.06$), driven by a slightly larger effect of the biasing context at intermediate (ambiguous) continuum steps. Finally, we observed a significant Step × Talker interaction; while differences between the two talkers were minimal at most continuum steps, there was a slight difference at Step 2, with participants more likely to make a "shine" response for the female talker (mean: 0.53, *SE*: 0.02) than for the
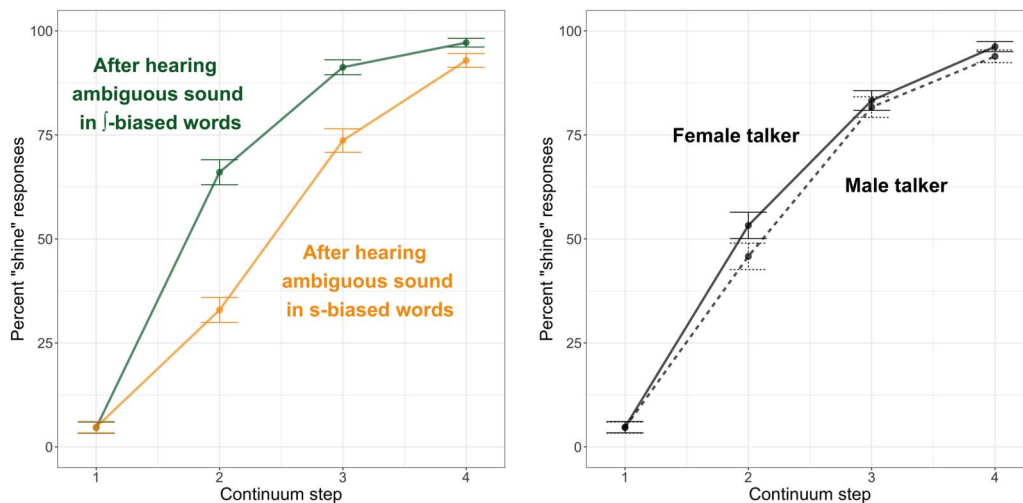

Figure 3. Phonetic categorization data from Experiment 1 (behavioral pilot). In both plots, the *x*-axis indicates the continuum step on the sign–shine continuum, while the *y*-axis indicates the percentage of "shine" responses at that step. Error bars indicate 95% confidence intervals around the mean. The left panel shows the data as a function of whether the talker previously produced ambiguous fricatives in /ʃ/-biased (green line) or /s/-biased (yellow line) contexts, and the right panel shows the data as a function of the talker's gender, female (solid line) or male (dashed line).

Table 1. Analysis of Experiment 1 phonetic categorization data (behavioral pilot)

Fixed effect	$\chi^2(1)$	p value	
Step	30.02	<0.0001	***
Bias	15.6	<0.0001	***
Talker	0.23	0.63	
Step × Bias	3.65	0.06	+
Step × Talker	5.54	0.02	*
Bias × Talker	0.57	0.45	
Step × Bias × Talker	0.72	0.40	

Note. *** indicates $p < 0.001$, ** indicates $p < 0.01$, * indicates $p < 0.05$, and + indicates $p < 0.10$.

male talker (mean: 0.46, *SE*: 0.02), regardless of whether the talker produced ambiguous fricatives in /s/-biased or /ʃ/-biased contexts during the exposure phase. No other effects were significant.

Discussion

In Experiment 1, listeners demonstrated talker-specific perceptual learning for two talkers (one female, one male); for one talker, an ambiguous fricative corresponded to the phoneme /s/, while for the other talker, the same ambiguous fricative corresponded to /ʃ/. Notably, talker-specific learning was observed with a relatively long session, in contrast to previous studies that have observed talker-specific learning in relatively short sessions (Kraljic & Samuel, 2007; Luthra et al., 2021). In a follow-up analysis, we tested whether the size of the learning (Bias) effect depended on which set of four runs (first, second, third, or fourth; see Figure 2) the participant was completing. Specifically, our statistical model tested for fixed effects of Set (scaled) in addition to the factors used in our initial analysis (Step, Bias, and Talker). We included each of the random effect terms used in the initial analysis as well their interactions with Set. We observed significant fixed effects of Step and Bias as well as a significant Step × Bias interaction. No other effects were significant. The lack of any significant interactions between Bias and Set suggests that the size of the learning effect was constant across the experimental session. Thus, Experiment 1 showed that listeners can form separate beliefs for how two talkers produce their speech sounds and can maintain these beliefs over a relatively long experimental session. Having observed robust learning effects in this experiment, we were able to conduct an fMRI experiment (Experiment 2) to probe the neural mechanisms through which listeners contact talker-specific beliefs for how acoustic-phonetic information maps onto phonetic categories.

EXPERIMENT 2

In Experiment 2, participants performed the same task as in Experiment 1 but did so in the scanner while fMRI data were collected. Thus, listeners were exposed to two talkers, one of whom produced an ambiguous fricative (a /s/-/ʃ/ blend) in place of a clear /s/ and one of whom produced the ambiguous fricative in lieu of /ʃ/; during exposure runs, phoneme identity was disambiguated through lexical information (e.g., *epi?ode*). After each exposure run, listeners performed a phonetic categorization task with a *sign–shine* continuum produced by the talker they had heard during the previous run. Critically, to resolve the phonetic identity of each test

stimulus (i.e., those presented in the phonetic categorization task), listeners must leverage talker information; this is particularly the case for the ambiguous stimuli, since for one talker an ambiguous fricative corresponds to /s/, but for the other it corresponds to /ʃ/. As such, a consideration of the functional activation during the phonetic categorization task can clarify the neurobiological mechanisms through which listeners condition phonetic identity on talker information.

First, we consider both the possibility that some set of regions represents both talker information and phonetic information. The extant literature has suggested that the posterior temporal cortex may play such a role, though the specific contributions of regions within posterior temporal cortex (e.g., STG, STS, MTG) remain unclear, as do the specific contributions of the right and left hemispheres (Evans & Davis, 2015; Formisano et al., 2008; Luthra et al., 2020; Myers & Mesite, 2014). The approach of the current study was to conduct a series of searchlight analyses, allowing us to identify voxels that are important for classification based on talker information as well as voxels that contribute to phonetic classification; of interest is the overlap between these sets of voxels, as these regions potentially constitute an integration site for talker information and phonetic information.

A core methodological feature of the current study is the use of searchlight analyses (Kriegeskorte et al., 2006). In this approach, classification analyses are performed within a roving "searchlight," and performance within a searchlight is assigned to the central voxel. Thus, searchlight analyses only consider the pattern of activation within spatially adjacent voxels. This is in contrast to previous work by Formisano et al. (2008), who also attempted to classify functional activation patterns based on talker and phonetic information but whose approach involved the use of recursive feature elimination (RFE; De Martino et al., 2008). Specifically, Formisano et al. trained their classifier on the patterns of activation across all the voxels in a prescribed region (the bilateral temporal cortex) and then used RFE to identify which voxels contributed most to classification. RFE can be particularly useful when the patterns of interest are distributed across a broad set of regions.

One potential concern relates to the spatial specificity of searchlight analyses—because performance within a searchlight is assigned to the central voxel, classification is undoubtedly also supported by surrounding voxels. Similarly, in RFE, because the machine learning algorithm has access to a relatively large set of voxels for performing classification, it is not necessarily the case that accurate classification can be performed using only the most discriminative voxels (as identified through RFE). That is, even if RFE can identify the set of voxels that are most informative for classification, voxels outside of this most discriminative set may still be necessary for accurate classification. For both types of MVPA, it is critical to consider the contributions of a voxel with respect to the full set of voxels under consideration. Because searchlight analyses consider only the local pattern of activation (as opposed to RFE, where the contributions of an identified voxel may be influenced by nonlocal voxels) and because a primary aim of the current study was to more precisely characterize the contributions of particular subregions in temporal cortex (STG, STS, and MTG), we opted to use a searchlight approach.

In addition to testing whether the integration of talker and phonetic information is achieved in a single region, we also considered the possibility that this integration is achieved through the coordinated activity of multiple regions. That is, there may be one set of regions that represents talker information and a distinct set of regions that represents phonetic information, and the two sets of regions may work together to achieve the integration of talker and phonetic detail. Since talker processing tends to rely relatively strongly on the right hemisphere and

phonetic processing tends to rely more strongly on the left hemisphere, this would likely be achieved specifically through functional connections between the left and right hemisphere. To assess this possibility, we first identified which voxels in the right hemisphere are sensitive to talker information (using a searchlight analysis, as above); we then tested whether the activity of these voxels predicts the activity of other brain regions during the phonetic classification task. Of interest was whether this latter connectivity analysis would identify left hemisphere regions implicated in phonetic processing—for instance, those in the left temporal cortex (e.g., STG and MTG; Blumstein et al., 2005; Desai et al., 2008; Scott et al., 2000), left parietal cortex (e.g., the supramarginal gyrus and angular gyrus; Blumstein et al., 2005; Lee et al., 2012; Raizada & Poldrack, 2007) or left frontal cortex (especially the left inferior frontal gyrus, or IFG; Blumstein et al., 2005; Lee et al., 2012; Rogers & Davis, 2017; Xie & Myers, 2018).

Experimental Design

Stimuli

Experiment 2 used the same stimuli as in Experiment 1.

Participants

Twenty individuals (14 female, 6 male; mean age: 21, age range: 18–28) were recruited from the University of Connecticut community. This sample size is consistent with previous MVPA studies (Correia et al., 2015; Feng et al., 2019; Joanisse et al., 2007; Kilian-Hütten et al., 2011; Lee et al., 2012; Luthra et al., 2020), which typically have an average sample size of about 17 (range: 10–30). All participants were right-handed native speakers of North American English and indicated that this was the only language they spoke prior to age 13. Participants did not have any hearing deficits, had normal or corrected-to-normal vision, had no history of neurological impairment, and met all MRI safety requirements (e.g., no ferromagnetic material in their bodies). Participants provided informed consent, and each participant was paid $50 for their time. All participants met the data inclusion criteria used in Experiment 1 (responding to more than 90% of the trials in both the exposure and test runs, showing at least 70% accuracy in phonetic classification of the unambiguous endpoints). All procedures were approved by the University of Connecticut Institutional Review Board.

Procedure

While the procedure of Experiment 2 was essentially the same as that of Experiment 1, slight timing changes were made to the design to make the experiment compatible with fMRI. As in Experiment 1, participants alternated between exposure (talker identification) runs and test (phonetic categorization) runs, completing 32 trials (16 ambiguous, 16 unambiguous) in each exposure run and 48 trials (12 repetitions at each of the four continuum steps) in each test run. Auditory stimuli were presented in silent intervals between scans (Edmister et al., 1999). As shown in Figure 4, each trial consisted of a 1 s "scan off" period, during which the auditory stimulus was presented, with the onset of the stimulus falling 200 ms into the silent interval; functional images were collected during a subsequent 2 s scan, during which a fixation cross was displayed on the screen. To appropriately model the hemodynamic response in each condition, each run included an equal number of "silent" trials as critical trials. Silent trials consist of a period of timing identical to a critical trial, except no stimulus is presented; these silent trials were interspersed between critical trials, thereby allowing us to jitter the onsets of critical trials. The specific order of these trials was determined using the optseq2 program (Greve, 2002). The response mappings were shown on the screen for the first 1,500 ms of each critical trial, and a fixation cross appeared on the screen during the second half of each critical trial as
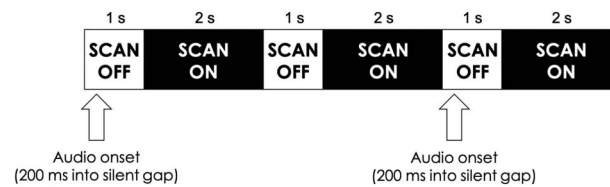
Figure 4. Clustered volume acquisition design for Experiment 2. On critical trials, auditory stimuli were presented in 1 s gaps between each 2 s scan period. Silent trials were interspersed between critical trials, allowing us to jitter the onsets of critical trials and thereby model the hemodynamic response appropriately.

well as throughout each silent trial. Experiment 2 was programmed using the OpenSesame experiment builder (Mathôt et al., 2012).

MRI data were collected at the University of Connecticut Brain Imaging Research Center using a 3T Siemens Prisma scanner (Siemens Medical Solutions, 2022) with a 64-channel head coil. Anatomical images were acquired using a T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) sequence (TR = 2,400 ms, TE = 2.15 ms, FOV = 256 mm, flip angle = 8°) with 1 mm sagittal slices. Axial-oblique functional echo planar images were acquired using a T2*-weighted sequence in ascending, interleaved order (TR = 3.0 s [effective TR of 2.0 s with a 1.0 s delay], TE = 25 ms, 52 slices, 2.5 mm thickness, in-plane resolution = 2 mm × 2 mm, FOV = 220 mm, flip angle = 62°, multiband acceleration factor = 2). In total, 64 volumes were acquired for each exposure run, and 96 volumes were acquired for each test run. The entire MRI session lasted approximately 90 min.

Stimuli were presented using a transducer system (Silent Scan SS-3100 made by Avotec (n.d.)) coupled to a pair of insert headphones (Conformal Headset made by Avotec (n.d.)). With this system, auditory stimuli can be delivered through two flexible tubes that penetrate foam ear tips. The ear tips are rolled down before they are placed in the participant's ear canal, where they then expand. In this way, the insert headphones are intended to provide hearing protection from the scanner noise while also allowing auditory stimuli to be delivered to the participant.

In-scanner Behavior

Behavioral results

Behavioral data were analyzed following the same approach as in Experiment 1. We used the same model structure as in Experiment 1; as before, the maximal random effects structure was also the most parsimonious.

Phonetic categorization data are plotted in Figure 5. In contrast to the results of Experiment 1, there appears to be a minimal effect of Bias in our fMRI participants, with only a slight difference in the categorization functions of listeners who heard the ambiguous fricatives in /ʃ/-biased contexts compared to /s/-biased frames (left panel). However, listeners appeared to differ dramatically in how they categorized stimuli produced by the female talker as compared to the male talker (right panel).

Results of the statistical analysis are provided in Table 2. We observed the expected effect of Step, with participants making more "sh" responses for stimuli that were closer to the "shine" end of the two continua. However, in contrast to Experiment 1, we did not observe any significant effects of Bias (all $p \geq 0.15$). Instead, we observed a main effect of Talker, with participants making more "sh" responses for the female talker (mean: 0.74, *SE*: 0.01) than for the male talker (mean: 0.54, *SE*: 0.01). We also observed a significant Step × Talker
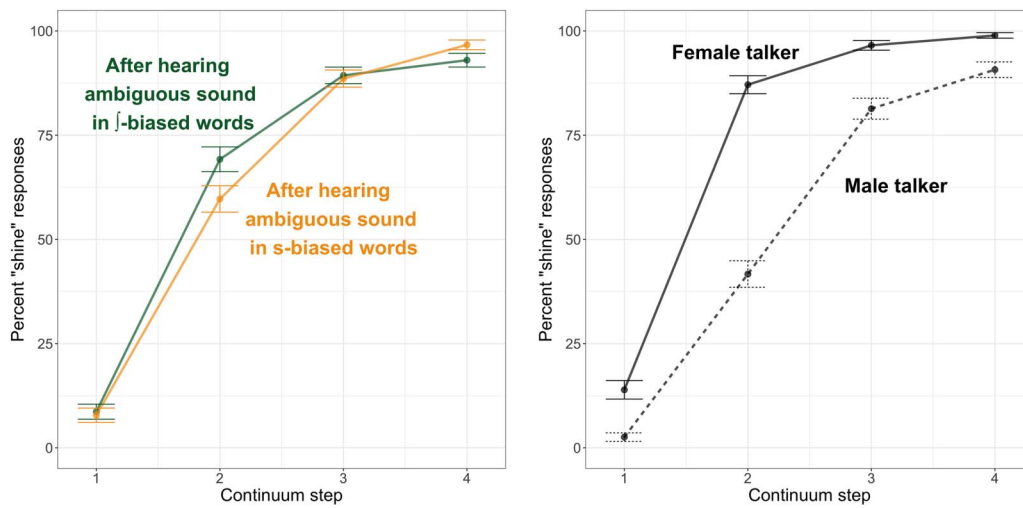
Figure 5. Phonetic categorization data from Experiment 2 (fMRI experiment). In both plots, the *x*-axis indicates the continuum step on the *sign–shine* continuum, while the *y*-axis indicates the percentage of "shine" responses at that step. Error bars indicate 95% confidence intervals around the mean. The left panel shows the data as a function of whether the talker previously produced ambiguous fricatives in /ʃ/-biased (green line) or /s/-biased (yellow line) contexts, and the right panel shows the data as a function of the talker's gender, female (solid line) or male (dashed line).

interaction, as there was a pronounced difference in how often participants made a "sh" response for the female talker (mean: 0.87, *SE*: 0.01) compared to the male talker (mean: 0.41, *SE*: 0.01) at Step 2 in particular.

Headphone Evaluation

One possible source of the discrepancy in the behavioral results of our two experiments might be the particular headphones that were used. Recall that because Experiment 1 was conducted online, participants used their own headphones, and a psychometric screener was used to ensure participants were using headphones instead of their computer loudspeakers. By contrast, stimuli for Experiment 2 were delivered through a pair of MRI-compatible insert headphones (with sound delivered via flexible tubes that penetrate foam earplugs). Note that because /s/ and /ʃ/ are distinguished primarily by spectral properties in relatively high frequency ranges, if key spectral information was not delivered faithfully to participants in Experiment 2,

Table 2. Analysis of Experiment 2 phonetic categorization data (fMRI experiment)

Fixed effect	$\chi^2(1)$	*p* value	
Step	37.05	<0.0001	***
Bias	2.1	0.15	
Talker	25.86	<0.0001	***
Step × Bias	0.76	0.38	
Step × Talker	5.38	0.02	*
Bias × Talker	0.22	0.64	
Step × Bias × Talker	0.46	0.5	

Note. *** indicates *p* < 0.001, ** indicates *p* < 0.01, * indicates *p* < 0.05, and + indicates *p* < 0.10.

then their categorization of our stimuli could have been affected. Thus, before presenting the MRI results of Experiment 2, we first consider the possibility that the headphone setup used for Experiment 2 may have contributed to the unexpected pattern of behavioral results.

There were two other minor differences in design between the two experiments, though we do not believe either of them would have contributed to the discrepancy in the pattern of results across the two experiments. In Experiment 2, stimuli were provided in silent intervals between periods of scanning. Though the scan sequence produces a considerable amount of auditory noise, the stimuli were presented 200 ms after the offset of the scanner noise, so we do not expect that perception of the speech stimuli would have been affected by energetic masking effects. We further note that in order to appropriately model the hemodynamic response, the timing of the trials in Experiment 2 was different from that of Experiment 1. A small number of studies have considered the role of timing in lexically guided perceptual learning, with research suggesting that disambiguating information needs to be encountered in advance of or shortly after a phonetically ambiguous segment (Jesse, 2021; Jesse & McQueen, 2011). However, to our knowledge, there is no evidence that the specific amount of time between trials can influence the degree or pattern of perceptual learning.

To directly assess the impact of the insert headphones, we coupled the insert headphones to an audio recorder and recorded the stimuli, allowing us to approximate the experience of participants in the scanner. We compared the spectral information in these recorded stimuli to the spectral information in the original stimuli.

Method

Stimuli were recorded in the MRI control room of the University of Connecticut Brain Imaging Research Center. Stimuli were presented using the same transducer system (the Avotec SS-3100 Silent Scan Audio System) and insert headphones (the Avotec Conformal Headset) as above. Recall that each insert headphone penetrates a foam ear tip. We trimmed a foam ear tip using a pair of scissors (taking care not to cut the headphone tubing itself), rolled the remaining foam, and inserted into in a Larson Davis AEC202 coupler (Larson Davis, 2021) for insert earphones, where the foam tip expanded to fill the coupler. The coupler was connected to a half-inch condenser microphone and preamplifier on a Larson Davis 824 sound level meter (Larson Davis, 2021). Finally, the sound level meter was connected to a Roland R-05 WAV recorder (Roland, 2022), using the AC-1/AC-2 output setting on the sound level meter.

Stimuli were analyzed using Praat (Boersma & Weenik, 2017). We used a spectral subtraction procedure (Boll, 1979), implemented through the "Remove Noise" function in Praat, to eliminate noise introduced by the recording procedure. Noise was defined as the spectral information present during a silent part of the recording (before stimuli began playing). Fricative onsets and offsets were marked visually by the first author. Spectral means (in Hz) were measured from the extracted fricative segments. To characterize the perceptual impact of frequency changes introduced by the headphones, we also report results on a perceptual (Mel) scale. Mel values were computed following the definition provided by Fant (1968):

$$Mel = \frac{1000}{log2}log\left(1 + \frac{frequency}{1000}\right)$$

Results

Spectrograms for a sample stimulus (a clear production of "sign" spoken by the female talker, taken from Step 1 of the female test continuum) are provided in Figure 6. The original stimulus
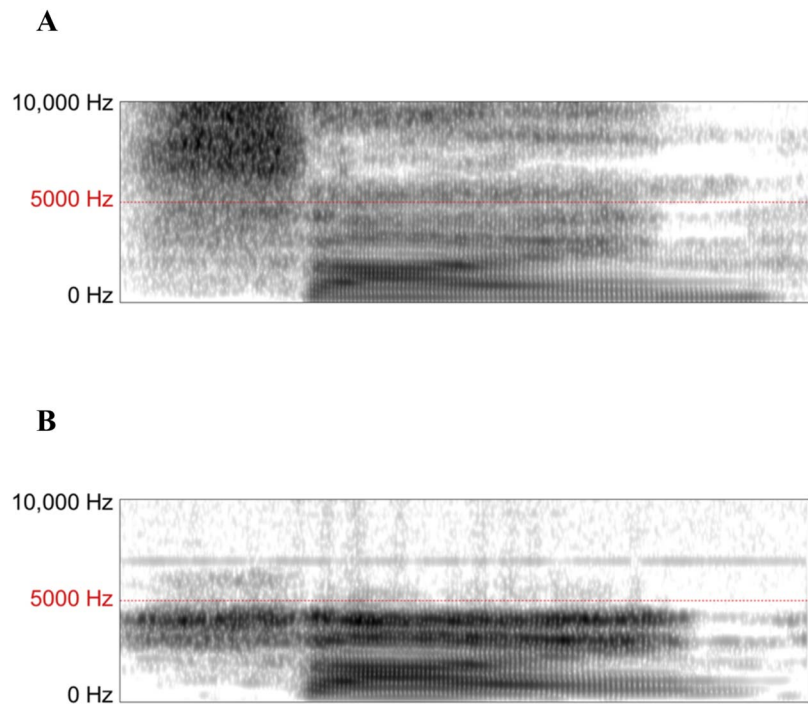
A

B

Figure 6. Sample spectrograms for a clear production of "sign" by our female talker (Step 1 of the *sign–shine* continuum). A. The spectrogram corresponds to the original stimulus; the spectral center of the fricative /s/ is 9048 Hz. B. The spectrogram corresponds to the stimulus recorded through the insert headphones used in the scanner and therefore approximates the spectral information delivered to our MRI participants. Frequencies above 5000 Hz are heavily attenuated, and the new spectral center of the /s/ fricative is 3950 Hz.

(Figure 6A) is characterized by strong amplitudes in the ~9000 Hz range during the period of frication. However, these frequencies were likely attenuated for our in-scanner participants, as shown in the spectrogram of the stimulus recorded from the insert headphones (Figure 6B).

The spectral properties of the stimuli are illustrated in Figure 7 (exposure stimuli in Figure 7A, test stimuli in Figure 7B). Figure 7A highlights spectral properties that were critical for driving learning in each group. For instance, for listeners in the /s/-biased group, lexically guided perceptual learning depends on a distinction between ambiguous /s/ segments and unambiguous /ʃ/ segments. While the spectral means for these stimuli were relatively well separated in the original stimuli (7329 Hz vs. 6405 Hz for the female talker, 5593 Hz vs. 4810 Hz for the male talker), the filtering introduced by the insert headphones meant that the spectral means were much closer together (3771 Hz vs. 3846 Hz for the female talker, 3684 Hz vs. 3553 Hz for the male talker).

Because learning depends on an acoustic difference between /s/ and /ʃ/ segments, we conducted two-sample one-tailed *t* tests on the spectral means of the /s/ and /ʃ/ stimuli presented during exposure, separately testing each talker and each biasing condition. That is, for each of the eight panels shown in Figure 7A, we tested whether the /s/ stimuli (red) had significantly higher spectral means than the /ʃ/ stimuli (blue). We applied a Bonferroni correction for multiple comparisons (alpha = 0.05 / 8 = 0.00625). Results are summarized in Table 3 and shown visually in Figure 7A. For the original stimuli, we found that the spectral means of the

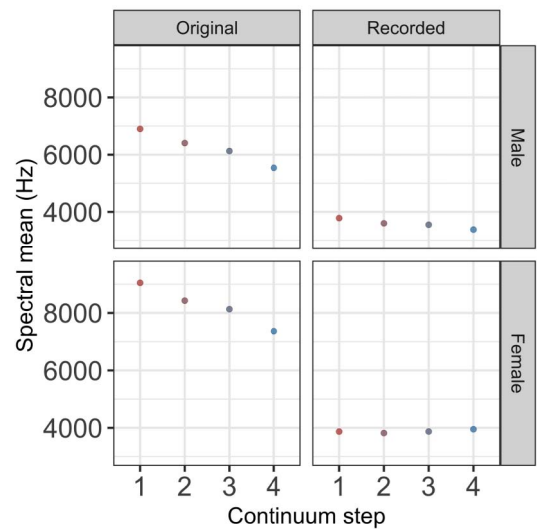**A Exposure stimuli**
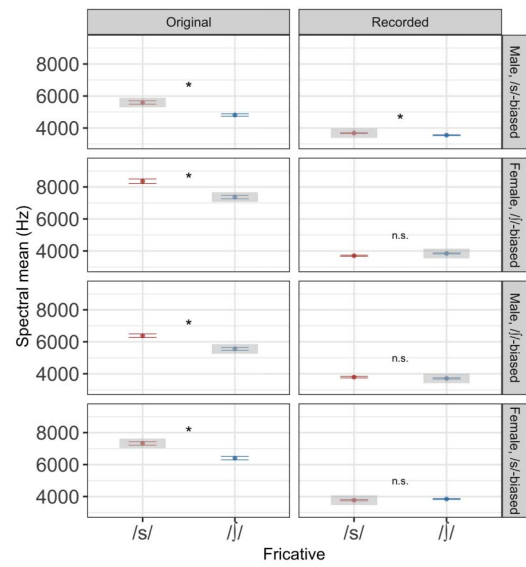
B Test stimuli

Figure 7. Spectral centers of gravity (in Hz) for the fricatives presented during exposure and test blocks. To assess the impact of the insert headphones used in the scanner, we compared the spectral properties of the original stimuli to those of the stimuli recorded through the insert headphones used in the scanner. Critically, /s/ should be associated with a higher spectral mean than /ʃ/. A. During exposure, one group of participants heard a male talker producing an ambiguous fricative in /s/-biased contexts and a female talker producing an ambiguous fricative in /ʃ/-biased contexts (top two rows). The other group of participants heard the male talker producing ambiguous fricatives in /ʃ/-biased contexts and the female talker producing ambiguous fricatives in /s/-biased contexts (bottom two rows). The /s/ fricatives are shown in red and /ʃ/ in blue. Gray boxes highlight the ambiguous fricative, the identity of which was determined with lexical information. Error bars indicate standard error. A significant difference between the spectral means of the /s/ and /ʃ/ stimuli is indicated with an asterisk (*), and a nonsignificant difference is indicated with "n.s." B. During test, listeners heard a continuum from "sign" (red points) to "shine" (blues); intermediate colors indicate ambiguous continuum steps. Rows indicate talker.

/s/ segments were significantly greater than those of the /ʃ/ segments. However, the insert headphones attenuated this difference: In three out of four cases, the /s/ segments did not have higher spectral means relative to the /ʃ/ segments for the stimuli recorded from the headphones.

Table 3. Results of one-tailed two-sample *t* tests evaluating whether the spectral means were higher for the /s/ stimuli than for the /ʃ/ stimuli for each combination of Talker and Bias

Talker/Bias	Original	Recorded
Male, /s/-biased	$t(30) = 6.27$, $p < 0.0001$, *	$t(30) = 3.39$, $p = 0.0010$, *
Female, /ʃ/-biased	$t(30) = 5.56$, $p < 0.0001$, *	$t(30) = -2.93$, $p = 0.9968$, n.s.
Male, /ʃ/-biased	$t(30) = 5.74$, $p < 0.0001$, *	$t(30) = 1.26$, $p = 0.1081$, n.s.
Female, /s/-biased	$t(30) = 6.09$, $p < 0.001$, *	$t(30) = -1.46$, $p = 0.9224$, n.s.

Note. * indicates significance at an alpha of $0.05/8 = 0.00625$ (Bonferroni correction for multiple comparisons). Results indicate that the acoustic difference between /s/ and /ʃ/—a key prerequisite for lexically guided perceptual learning—was disrupted by the insert headphones used in the scanner.

Discussion

To better approximate the listening environment of participants in the scanner, we recorded our stimuli as they were played out through the insert headphones and performed an acoustic analysis. We found that frequencies above 5000–6000 Hz were dramatically attenuated by the insert headphones. We suspect that this is the primary cause for the discrepancy in the behavioral data between Experiment 1 and Experiment 2. While participants in Experiment 2 reported hearing instances of /s/ and /ʃ/, the attenuation of key spectral information inhibited lexically guided perceptual learning. In our analysis of the spectral centers of gravity, we found that the acoustic difference between the /s/ and /ʃ/ segments (a key ingredient for learning) was disrupted by the headphones.

Recall that the primary aim of this study was to probe the neural systems that allow listeners to access talker-specific generative models and that we were specifically interested in generative models that had been shaped by lexically guided perceptual learning. Because of the way our headphones impacted stimulus delivery, listeners in Experiment 2 were not influenced by the lexically biasing information in the exposure runs. Nevertheless, listeners did show differential phonetic categorization profiles for the two talkers. As such, Experiment 2 can still allow for a fruitful examination of the neural systems supporting access to talker-specific beliefs; we return to this issue in the General Discussion.

MRI Preprocessing

Preprocessing of the fMRI data was achieved in AFNI (Cox, 1996), with slightly different processing pipelines used for the functional connectivity analyses and the searchlight analyses. For the functional connectivity analyses, oblique functional images were first warped to a cardinal orientation and resampled to isotropic (2 mm × 2 mm × 2 mm) voxels using the 3dWarp command. Next, functional data were registered to the participant's anatomical data, aligned to the first volume of each functional run to correct for within-run motion using a six-parameter rigid body transform, and warped from the participant's native space to Talairach space (Talairach & Tournoux, 1988); these three affine transformations were applied simultaneously to the data to minimize interpolation penalties and were implemented using the align_epi_anat.py program built into afni_proc.py. Functional data were then smoothed using a 4 mm full-width half-maximum kernel and scaled to have a mean value of 100. Two participants only completed 12 of the 16 functional runs, and for one of these participants, substantial motion after the sixth run necessitated a second anatomical scan. For the latter participant, the first half of the functional data was registered to the first anatomical data set and the second

half was registered to the second anatomical data set, but the preprocessing pipeline was otherwise identical to the one used for the other participants. Following preprocessing, we also conducted univariate analyses of the data (see Supporting Information available at https://doi .org/10.1162/nol_a_00091).

For searchlight analyses, preprocessing involved transforming functional data from an oblique orientation to a cardinal one, registering functional data to the anatomical data, and aligning the functional data to the first volume of the corresponding run. Consistent with the univariate analysis, registration and alignment were performed using a single transformation. Note that for these analyses, data were not transformed to Talairach space, such that searchlight analyses for each participant could be conducted in their native space. However, as described above, one participant moved substantially midway through the experiment, necessitating the collection of a second anatomical scan. For that participant, the 3dFractionize command was used to transform the Talairach-warped functional images (generated as part of preprocessing for the functional connectivity analyses) back to the participant's native space; the first anatomical scan (in native space) was used as a template for this transformation. In this way, all functional images for this participant were in the participant's native space and registered to the same anatomical image.

Searchlight Analyses

Methods

Feature estimation.   After preprocessing, we estimated voxel-wise beta weights for every phonetic categorization trial, using a least-sum-squares regression approach suited for fast event-related fMRI designs (Mumford et al., 2012). In this approach, a separate regression analysis is performed for every trial, with each regression involving a critical regressor for the trial of interest and a nuisance regressor for all other trials in that condition (here, all other instances of that same continuum step spoken by that same talker). This feature estimation step was implemented using the -stim_times_IM flag in the 3dDeconvolve command and the 3dLSS command.

Talker classification.   The goal of our first analysis was to identify voxels where the local pattern of activation was informative of talker identity. To this end, a support vector machine was trained on the patterns of activation within an 8 voxel (16 mm) radius searchlight, with activation patterns labeled with respect to talker identity (i.e., female or male). Multivoxel pattern analyses were implemented using The Decoding Toolbox (Hebart et al., 2015). We decided a priori to limit our analyses to anatomical regions that are broadly associated with speech processing—namely, the inferior frontal, middle frontal, insular, superior temporal, transverse temporal, middle temporal, supramarginal, and angular cortices bilaterally, as defined in the AFNI Talairach atlas. This group mask is visualized in Figure 8. We further limited our analyses
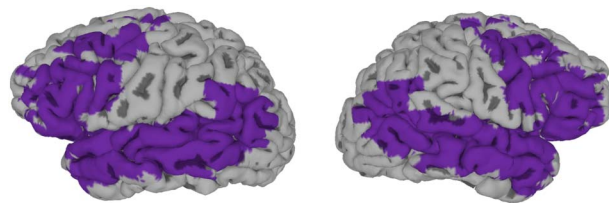

Figure 8.   Voxels considered in analyses. For all analyses, we only considered voxels in left-hemisphere regions that have been implicated in language processing as well as the corresponding regions in the right hemisphere. For visualization purposes, volumetric clusters were projected to a surface reconstruction using FreeSurfer (Fischl, 2012) and SUMA (Saad & Reynolds, 2012).

to voxels that had been imaged in all 20 participants; to achieve this, our group mask was transformed to each subject's native space using the 3dFractionize command in AFNI (Cox, 1996).

Recall that during the experiment, runs were organized into sets of four, with each set involving an exposure run and a phonetic categorization run for each talker (Figure 2). Cross validation was achieved using a leave-one-set-out approach, such that in each cross-validation fold, the classifier was trained on data from three phonetic categorization runs from each talker (or, for the two participants who did not complete the experiment, on two phonetic categorization runs for each talker) and tested on data from two held-out runs, one for each talker. In this way, there was an equal number of male and female training trials for each cross-validation fold. For each searchlight, the classification performance score at test (accuracy minus chance) was assigned to the central voxel of the searchlight. The support vector machine's regularization parameter was set at C = 1, and default settings were used for all other parameters.

Subject-level classification maps were converted to Talairach space and blurred using a 4 mm full-width half-maximum kernel in AFNI. The resultant classification maps were then submitted to a one-sample *t* test using the 3dttest++ command. To correct for multiple comparisons, we used a cluster size threshold of 116 voxels (voxel-wise $p < 0.05$, cluster-level $\alpha < 0.05$).

Phonetic classification in regions associated with speech perception. In a second searchlight analysis, we sought to identify voxels where the local activation pattern was informative of phonetic identity. For this analysis, we considered the same set of voxels that were used in the talker classification searchlight analysis. This analysis considered only trials in which participants heard the unambiguous continuum endpoints, and trials were labeled with respect to the expected response (i.e., Step 1 labeled as /s/ and Step 4 labeled as /ʃ/). Thus, there was an equal number of trials in each training class. As before, we used an 8 voxel radius searchlight with a leave-one-set-out cross-validation scheme and the same support vector machine parameters as before. By-participant classification maps were converted to Talairach space, blurred using a 4 mm full-width half-maximum kernel, and submitted to a one-sample *t* test, and the same cluster correction was applied to the resultant group map. Of interest is the extent to which this phonetic classification searchlight analysis identified the same voxels as the talker classification searchlight analysis described above.

Phonetic classification in regions that support talker classification. Our final searchlight analysis also sought to identify voxels that were informative of phonetic identity, but in this analysis, we limited our search to the voxels that were identified as showing above-chance accuracy in the talker classification searchlight analysis (rather than testing the full set of regions involved in speech processing, as above). Because these ROIs were defined at the group level, we transformed the functional ROIs to each subject's native space using the 3dFractionize command in AFNI. For this analysis, we used a smaller (3 voxel) radius searchlight, since the searchlight analysis was being conducted over a smaller set of voxels than before; this also improves the spatial resolution of the analysis, as classification in an individual voxel is supported by a relatively small number of neighboring voxels.

In contrast to the previous phonetic classification searchlight analysis, this analysis considered all phonetic categorization trials (i.e., both ambiguous and unambiguous continuum steps), and activation patterns were labeled based on the participants' trial-by-trial behavioral

responses. To ensure that the classifier was not systematically biased toward whichever response (/s/ or /ʃ/) a participant made more often, we balanced the number of training trials from each class by subsampling from the more represented class; this balancing procedure was repeated 10 times for each cross-validation fold, and the average of the 10 repetitions was taken as the classifier's performance for that cross-validation fold. Note that this approach would have been computationally impractical for the previous phonetic classification searchlight analysis, as it would have involved performing 10 repetitions for each cross-validation fold for a relatively large number of voxels (i.e., for all the voxels shown in Figure 8). Here, this approach is tenable because of the relatively small ROIs constraining the searchlight analysis. As before, the support vector machine's regularization parameter was set at $C = 1$, and default settings were used for all other parameters. The resultant subject-wise classification maps were warped to Talairach space, blurred using a 4 mm full-width half-maximum kernel, and submitted to a one-sample t test. Because a cluster correction had been applied to define the ROIs for this analysis, no further cluster correction was applied at this point.

Results

Results of the searchlight analyses are provided in Table 4 and visualized in Figure 9. As shown in Figure 9A, we found that our classifier achieved above-chance classification of talker

Table 4.   Results of the searchlight analyses of fMRI data (voxel-wise $p < 0.05$, cluster-level $\alpha < 0.05$)

Anatomical region	Maximum intensity coordinates			Number of activated voxels	t value / z value
	x	y	z		
Talker classification in regions that support speech perception					
Right superior temporal gyrus / right superior temporal sulcus / right middle temporal gyrus	55	−25	8	1,005	4
Left insula / left Heschl's gyrus	−45	−21	20	162	2.8
Phonetic classification in regions that support speech perception					
Left superior temporal gyrus / left middle temporal gyrus / left insula	−53	−39	14	3,624	4.8
Left inferior frontal gyrus / left middle frontal gyrus	−47	13	42	1,508	3.7
Right inferior frontal gyrus / right middle frontal gyrus	43	13	34	1,506	4
Right superior temporal gyrus / right superior temporal sulcus / right middle temporal gyrus	55	−49	2	1,274	3.2
Left middle frontal gyrus	−35	49	−2	370	3.6
Overlap of Talker classification and Phonetic classification maps					
Right superior temporal sulcus	53	−37	−2	67	
Left Heschl's gyrus / left insula	−41	−25	12	62	
Phonetic classification in regions that support talker classification					
Right superior temporal sulcus	53	−37	−2	19	3.1

A Talker classification
in regions that support speech perception

B Phonetic classification of
unambiguous tokens
in regions that support speech perception

$t = 2.1$ $F = 4.1$

$t = 2.1$ $t = 5.8$

C Overlap of (A) and (B)

D Phonetic classification of
ambiguous and unambiguous tokens
in regions that support talker classification

$z = 2.0$ $z = 3.3$
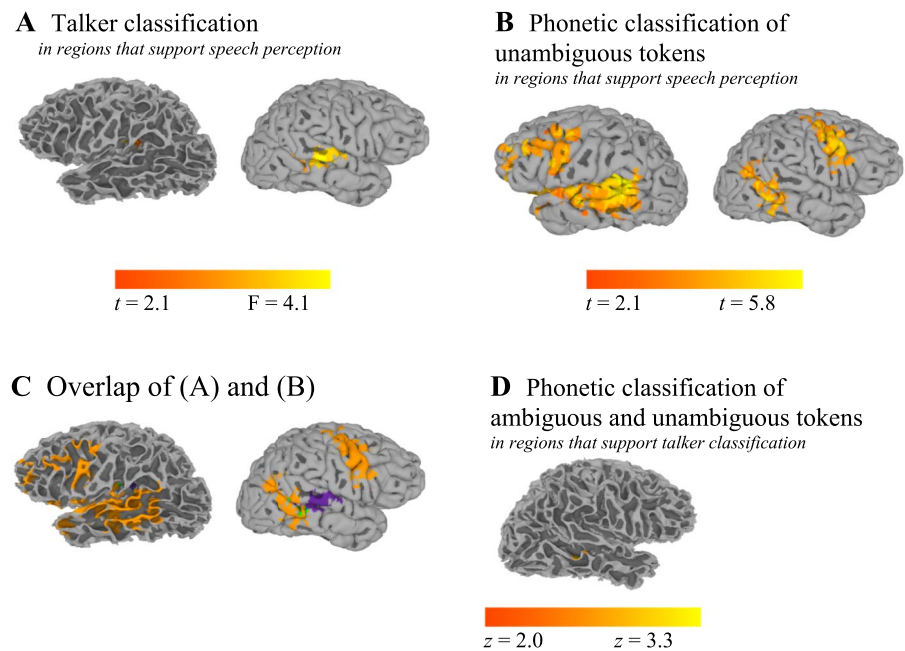
Figure 9. Brain regions identified using searchlight analyses. A. Set of regions sensitive to talker identity. B. Set of regions sensitive to phonetic identity. C. These two sets of regions overlapped, suggesting that some regions (left Heschl's gyrus/insula, right superior temporal sulcus; shown in green) are sensitive both to talker information and phonetic identity. Voxels colored in purple are sensitive to talker information and voxels colored in orange are sensitive to phonetic information. D. We also conducted a searchlight analysis to identify voxels that were sensitive to phonetic identity, but for this analysis, we limited the set of candidate voxels to those identified in our talker classification analysis (panel A). By default, we show the pial surface of the brain, though for maps in which our clusters extended into sulci, we show the white matter surfaces instead. Alternate views of Figure 9D are provided in the Supporting Information.

identity when considering only the local activation patterns in the right STG/STS (extending partly into the right MTG) as well as in early auditory cortex in the left hemisphere (left Heschl's gyrus). A support vector machine trained to make classifications on the basis of phonetic identity performed above chance when given local activation patterns in the superior/middle temporal cortex or inferior/middle frontal cortex, whether on the left or on the right (Figure 9B). Figure 9C shows the overlap between these two classification maps; results suggest that the local patterns of activation in left primary auditory cortex or in the right STS can be used to classify trials both on the basis of talker identity (whether the listener heard the male or female talker) and on the basis of phonetic identity (whether the listener heard a clear /s/ or a clear /ʃ/).

In a separate searchlight analysis, we attempted to classify stimuli with regard to how they were interpreted on a trial-by-trial basis (regardless of the actual acoustics). For this analysis, we only considered voxels that had been identified by the talker classification searchlight analysis (Figure 9A). As shown in Figure 9D (as well as in Figure S2), we found that voxels in the right STS could be used both for classification of phonetic identity and of talker identity; no clusters in the left Heschl's gyrus ROI showed above-chance classification of both phonetic detail and talker. This analysis converges with the overlap analysis depicted in Figure 9C: Namely, voxels in right STS were sensitive to both talker and phonetic information. Taken together, the two sets of searchlight analyses suggest a role for the right STS in integrating talker information and phonetic identity.

Table 5. Functional connectivity results (right hemisphere seed, voxel-wise *p* < 0.05, cluster-level α < 0.05)

Anatomical region	Maximum intensity coordinates			Number of activated voxels	*t* value
	x	*y*	*z*		
Right middle frontal gyrus	53	9	−34	692	2.2
Right middle temporal gyrus / right superior temporal gyrus	67	−27	−4	312	3.2
Left middle frontal gyrus / left inferior frontal gyrus	−53	9	−34	274	3
Left superior temporal gyrus / left middle temporal gyrus	−63	−13	−8	255	2.1

Note. The seed region was a right hemisphere cluster shown to support talker classification.

Functional connectivity analyses

Methods

To examine the extent to which the integration of talker information and phonetic information is achieved through interactions among multiple brain regions, we conducted a set of functional connectivity analyses. Each analysis used one of two seed regions identified through the talker classification searchlight analysis (Figure 9A)—that is, one analysis used the right STG/STS/MTG cluster as a seed, and the other used the left insula/left Heschl's gyrus cluster as a seed region.

For each seed region, we conducted a regression analysis including several predictors. First, we included the smoothed, scaled time course of our seed region; this represents the baseline activity of the seed region. The second predictor was the convolution of the time course of the seed region with a vector of the onset times for all the phonetic categorization trials; this predictor therefore represents the activity of the seed region during the phonetic categorization trials (above and beyond its baseline activity) and is the regressor of main interest for these analyses. We also considered some regressors of no interest. First, stimulus onset times for each condition (ambiguous exposure trials, unambiguous exposure trials and each of the four
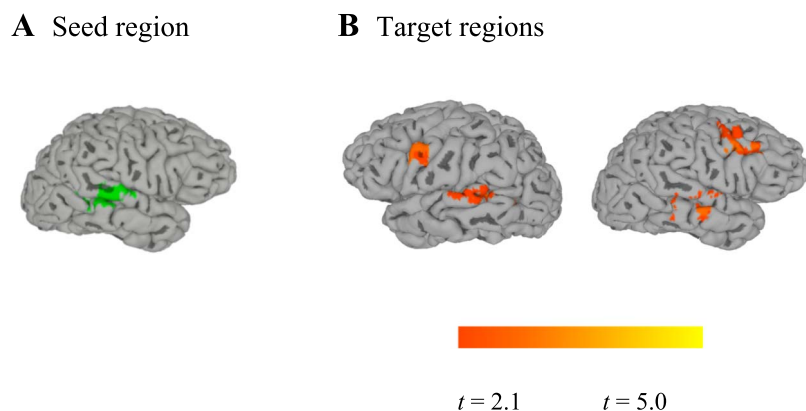
A Seed region   **B** Target regions

t = 2.1   *t* = 5.0

Figure 10. Functional connectivity analysis to investigate how the integration of phonetic information and talker information might be achieved through the coordinated activity across multiple brain regions. A. Our seed region (green) was functionally defined as the set of right hemisphere voxels that supported talker classification, as identified by our searchlight analysis. B. Target regions identified in this analysis showed increased functional connectivity with the seed region during phonetic categorization trials.

Table 6. Functional connectivity results (left hemisphere seed, voxel-wise $p < 0.05$, cluster-level $\alpha < 0.05$)

Anatomical region	Maximum intensity coordinates			Number of activated voxels	t value
	x	y	z		
Left superior temporal gyrus / left middle temporal gyrus	−63	−11	6	943	8.3
Right middle temporal gyrus / right superior temporal gyrus / right insula	67	−27	−4	902	6.6

Note. Seed region was a left hemisphere cluster shown to support talker classification.

continuum steps presented in the phonetic categorization task, modeled separately for each talker) were convolved with a gamma function to generated idealized hemodynamic response functions. Finally, each regression also included a third-order polynomial term (to account for scanner drift over the course of the run) as well as the six motion parameters estimated during the alignment step of preprocessing.

To identify which regions were functionally connected with each seed region during the phonetic categorization task, we conducted a one-sample t test on the subject-wise beta estimates for the Seed × Phonetic Categorization task regressor using the 3dttest++ command in AFNI (Cox, 1996); as in the searchlight analyses, results were masked to only include the voxels shown in Figure 8. To correct for multiple comparisons, Monte Carlo simulations were performed on the group mask using the 3dClustSim command, using average smoothness values estimated by applying the 3dFWHMx command to the residual time series data from each regression. In this way, we determined that we would need a minimum of 209 voxels for a statistically significant cluster (voxel-wise $p < 0.05$, cluster-level $\alpha < 0.05$).

Results

In one functional connectivity analysis, we examined which other brain regions were functionally connected with the right STG/STS seed during the phonetic categorization task. As described in Table 5 and illustrated in Figure 10, we found that on phonetic categorization
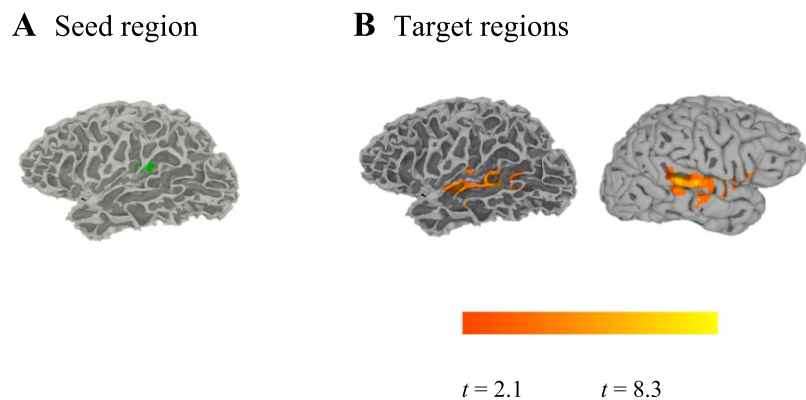
A Seed region   **B** Target regions

$t = 2.1$ $t = 8.3$

Figure 11. Functional connectivity analysis to investigate how the integration of phonetic information and talker information might be achieved through the coordinated activity across multiple brain regions. A. Our seed region (green) was functionally defined as the set of left hemisphere voxels that supported talker classification, as identified by our searchlight analysis. B. Target regions identified in this analysis showed increased functional connectivity with the seed region during phonetic categorization trials.

trials, the activity of the right STG/STS showed significant correlations (relative to the baseline degree of functional connections) with the activity of the right MFG, the right MTG/STG, the left IFG, and the left posterior STG.

In our second functional connectivity analysis, we examined which brain regions were functionally connected with the left insula/Heschl's gyrus seed region that had been identified by the searchlight analysis. As described in Table 6 and illustrated in Figure 11, we found that on phonetic categorization trials, the activity of the left insula/Heschl's gyrus showed significant correlations (relative to the baseline degree of functional connections) with the activity of the left STG/MTG and the right STG/MTG/insula.

GENERAL DISCUSSION

In this study, listeners were exposed to two talkers who produced an ambiguous fricative ("*?*") either in place of /s/ or in place of /ʃ/. During exposure runs, ambiguous fricatives were encountered in disambiguating lexical frames (e.g., *epi?ode*). During test phonetic categorization runs, listeners heard these sounds in the absence of lexical context (e.g., *a?i*) and had to condition their interpretation of stimuli on acoustic details correlated with talker identity. Of interest were the patterns of functional activation elicited during the test runs. Results of searchlight analyses indicate that the patterns of activation in the right STS contain information about talker identity as well as about phonetic identity, suggesting that the right STS may serve as an interface between a left-lateralized phonetic processing system and a right-lateralized neural system for talker processing. Functional connectivity analyses examined how the activity of other regions related to the activity of the right STS when listeners made phonetic categorization decisions. Notably, we observed increased functional connectivity between the right STS and right hemisphere regions that support talker processing, as well as increased functional connections between the right STS and regions in the left hemisphere implicated in phonetic processing.

Talker-Specific Phonetic Processing

Strikingly, the behavioral data elicited by our fMRI experiment (Experiment 2) did not comport with the results from our behavioral pilot (Experiment 1). In Experiment 1, listeners' performance during the phonetic categorization task depended on the contexts in which they had previously encountered the speech sounds (i.e., whether lexical knowledge biased them to interpret ambiguous sounds from that talker as /s/ or /ʃ/) but did not depend strongly on talker's gender (i.e., female or male). By contrast, listeners in Experiment 2 were not influenced by the lexically biasing information presenting during exposure runs but were strongly influenced by the talker's gender.

We believe that the discrepancy in the behavioral results is most likely attributable to the insert headphones that were used in the scanner. In a follow-up headphone evaluation, we recorded stimuli directly from the insert headphones and conducted an acoustic analysis on the recordings. We found that frequencies above the 5000–6000 Hz range were almost completely attenuated by the insert headphones. Our acoustic analysis also indicated that the spectral means in our stimuli—a key acoustic property for distinguishing between /s/ and /ʃ/—were altered by the insert headphones, and the attenuation affect was not equal for the two talkers. From a methodological perspective, the current work reveals that researchers must characterize the frequency response profile of headphones they are using (particularly for MRI studies) and evaluate the impact this will have on stimulus presentation. If potential

attenuation effects cannot be mitigated, it becomes critical to ensure during piloting that the behavioral effects of interest can still be elicited when, for example, high frequencies are attenuated. To do so, researchers might construct an auditory filter based on the frequency response profile of the MRI headphones, apply the filter to their auditory stimuli, and then conduct behavioral pilot tests to understand how the in-scanner headphones might influence the behavioral effect of interest. Researchers might also consider other (e.g., fiber-optic) headphone systems that can more faithfully deliver high frequencies, though some attenuation should still be expected, and appropriate pilot testing should still be conducted with other stimulus delivery setups. Researchers may also need to weigh potential attenuation of high frequencies against other important considerations (e.g., the quality of fit of the headphones, which will also influence the degree of scanner noise heard by the participant).

Listeners in Experiment 2 did show differential phonetic categorization profiles for the two talkers. However, these talker-specific beliefs were likely guided by acoustic details in the signal, rather than by lexical information specifically. Note that because the male talker's speech was generated by shifting the pitch and formant ratios from the female talker's speech, the acoustics of the two talkers' productions differ systematically; for instance, the spectral center of gravity differs between the two talkers, as shown in Figure 7. Thus, when listeners conditioned phonetic identity on talker information (whether categorical representations of talker identity or whether simply acoustic patterns correlated with talker identity), they may have done so based on the specific acoustic properties of the stimuli. Of interest, prior research has shown that listeners can condition phonetic identity on the perceived gender of the talker as well as other socio-indexical properties (Johnson et al., 1999; Kleinschmidt, 2019); it is therefore an empirical question whether listeners' beliefs in the current study were shaped by the raw acoustic information in the stimuli or by the acoustic details conditioned on gender (or some other socio-indexical variable). More generally, future work is needed to investigate how the neural basis of accessing talker-specific generative models may differ depending on the specific source of information (lexical, socio-indexical, acoustic) that shapes a listener's beliefs.

The Role of the Right Superior Temporal Sulcus

We conducted a series of searchlight analyses to investigate whether any brain regions simultaneously represented information about talker and about phonetic identity. An initial analysis found that talker identity could be recovered from local patterns of activation in the right posterior temporal cortex (primarily voxels in the STG and STS extending into the MTG) as well as from the patterns of activity in left Heschl's gyrus. To test if any of these patterns also contained information about phonetic identity, we used both a conjunction approach and a ROI analysis. For the conjunction analysis, we examined the overlap between the voxels identified in our talker classification analysis and voxels identified in a parallel phonetic classification analysis. We found overlap between these two maps in right STS as well as in left Heschl's gyrus. For the ROI analysis, we conducted a phonetic classification analysis within the right-hemisphere cluster (STG/STS/MTG) identified by the talker classification analysis. Here as well, we found that the local pattern of activation in the right STS could be used to perform both talker classification and phonetic classification, corroborating the results of the conjunction analysis. Overall, our results clarify the contributions of right STG, STS and MTG for processing talker information and phonetic information. In particular, they suggest that the right STG is a key region for representing talker information, that the right MTG is a key region for representing phonetic information, and that the right STS represents both sources of

information. Thus, the present work indicates that the right STS may serve as an integration site for talker information and phonetic information, allowing listeners to condition phonetic identity on talker information.

Our findings support and extend results from a previous MVPA study by Formisano et al. (2008), who also found that the right STS was important for decoding both phonetic identity and talker identity. However, we differentiate our study from theirs in four key ways. While Formisano et al. presented listeners with three Dutch vowels (/a/, /i/, and /u/) produced in isolation by three talkers, we presented our speech sounds (/s/ and /ʃ/) in lexical contexts, thereby increasing the ecological validity of the results. Secondly, we included phonetically ambiguous stimuli in our study, whereas Formisano et al. used only unambiguous stimuli; our goal in including phonetically ambiguous stimuli was to encourage reliance on talker information, which could be used to guide the interpretation of the phonetically ambiguous stimuli. Thirdly, the present work is distinguished from the previous study by the use of a searchlight-based approach, which allows us to make stronger claims about the spatial specificity of our results. This is because a searchlight analysis tests whether above-chance classification can be achieved on the basis of the local pattern of activation. By contrast, Formisano et al. considered a large set of voxels (bilateral auditory cortex) and used RFE to identify the most informative voxels. However, voxels that are identified through this approach must be considered in the context of the full set of voxels that supported classification (Hebart & Baker, 2018). Thus, even though their analyses implicated voxels in the right STS as representing both talker identity and phonetic identity, it is possible that a classifier given only those voxels would not be able to perform above chance. The convergence between our results and those from Formisano et al. is encouraging, given differences in materials and analysis approach. Finally, we designed the current study to investigate how listeners access their beliefs of how different talkers produce their speech sounds as those beliefs were updated through perceptual learning; that is, we wanted our listeners to learn the phonetic signature of each talker during the experiment. However, as discussed above, listeners in Experiment 2 did not show the expected learning effects behaviorally, a result we suspect was driven by the particular auditory equipment used in our study. Thus, in both the present work and the work by Formisano et al., listeners' phonetic processing likely reflects their prior assumptions of how the two talkers would produce their speech sounds rather than being driven by recently updated generative models.

While the present results suggest that the right STS plays an important role in allowing listeners to integrate phonetic information and talker information, future work will be needed to examine the generalizability of these results. For instance, previous work has shown that phonetic features such as voicing, manner, and place of articulation are distributed across the superior temporal lobe (Arsenault & Buchsbaum, 2015; Mesgarani et al., 2014), so it will be important to examine whether the right STS plays a similar role in conditioning phonetic identity on talker information for speech sound distinctions that rely on other featural differences. Additionally, it is unclear whether the right STS is *necessary* for integrating phonetic detail and talker information. Notably, our conjunction analysis suggested that phonetic information and talker information are also simultaneously represented in left Heschl's gyrus. Thus, even when listeners are unable to recruit the right STS (whether as a result of brain damage or due to the application of focal brain stimulation), listeners might still be able to adapt to talker-specific phonetic variation if they are able to recruit left Heschl's gyrus. More generally, an important future direction will be to delineate the differential contributions of the right STS and left Heschl's gyrus for contacting talker-specific generative models of how acoustics map onto speech sounds.

A Multi-System Approach to Talker-Specific Phonetic Processing

In addition to testing whether phonetic information and talker information are simultaneously encoded in the right hemisphere, we also tested the hypothesis that the process of conditioning phonetic identity on talker information is achieved through the coordinated activity of multiple brain regions. Such an investigation was motivated in part by previous work showing an increase in the functional connectivity of the left and right temporal cortices during speech processing (von Kriegstein et al., 2010). We found that on phonetic categorization trials, there was an increase in the connectivity of our right STG/STS seed region (defined functionally from our talker classification searchlight analysis) and several target regions: (1) the right MFG, a region that has been implicated for challenging vocal identity processing tasks (e.g., when listeners must hold vocal auditory objects in working memory; Stevens, 2004); (2) a slightly anterior portion of the right MTG, which is notable since anterior parts of the right temporal cortex have been linked to explicit vocal identification (see Luthra, 2021, for review), and (3) two left hemisphere regions thought to support phonetic processing: the left IFG (Lee et al., 2012; Myers, 2007; Myers et al., 2009; Rogers & Davis, 2017; Xie & Myers, 2018) and the left posterior STG (Desai et al., 2008; Liebenthal et al., 2010; Luthra et al., 2019; Mesgarani et al., 2014; Myers, 2007; Yi et al., 2019). Overall, then, when listeners performed a phonetic categorization task, there was an increase in the functional connections between: (1) a right STS region shown to represent talker information, (2) other right hemisphere regions that support vocal identity processing, and (3) left hemisphere regions implicated in phonetic processing. Taken together, these results suggest that the right STS may serve as an interface between the neural systems for talker processing and phonetic processing and that access to talker-specific generative models may be supported through the coordinated activity of these two systems. Consistent with this latter point, we also observed increased functional connectivity during the phonetic categorization task between our left Heschl's gyrus seed region (again, defined functionally from our talker classification searchlight analysis) and the STG/MTG bilaterally.

Conclusions

The speech signal simultaneously conveys information about who is talking and information about what is being said, and previous work suggests a substantial degree of interdependence between talker information and linguistic information (Maguinness et al., 2018; McGettigan & Scott, 2012; Mullennix & Pisoni, 1990). Several studies have demonstrated talker-specific effects in speech perception and in recognition memory (Kapnoula & Samuel, 2019; Luthra et al., 2018; Mattys & Liss, 2008; McLennan & Luce, 2005; Nygaard et al., 1994; Palmeri et al., 1993; Theodore et al., 2015), and theoretical accounts suggest that speech perception is guided by a listener's expectations of how a talker will produce their speech sounds (Crinnion et al., 2020; Kleinschmidt, 2019; McMurray & Jongman, 2011).

An emergent body of literature has suggested that the right hemisphere may play an important role in adapting to the idiosyncratic ways that different talkers produce their speech sounds (Luthra, 2021; Luthra et al., 2020; Myers & Mesite, 2014; Myers & Theodore, 2017). In the current work, we used searchlight analyses to show that talker information and phonetic identity are both represented in the local patterns of functional activation in the right STS. We also found evidence that the integration of talker information and phonetic information may depend on the coordinated activity of this right STS region, right hemisphere regions that support vocal identity processing, and left hemisphere regions that support phonetic processing. Though theoretical models of speech perception have suggested a relatively

minor role for the right hemisphere (at least in comparison to its left counterpart), the current study suggests that the right hemisphere meaningfully contributes to speech perception, particularly when listeners must adapt to talker-specific phonetic variation.

AUTHOR CONTRIBUTIONS

Sahil Luthra: Conceptualization; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing – original draft, Writing – review & editing. **James S. Magnuson**: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing – review & editing. **Emily B. Myers**: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing – review & editing.

REFERENCES

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, *115*(6), 3171–3183. https://doi.org/10.1121/1.1701898, PubMed: 15237841

Andics, A., McQueen, J. M., & Petersson, K. M. (2013). Mean-based neural coding of voices. *NeuroImage*, *79*, 351–360. https://doi.org/10.1016/j.neuroimage.2013.05.002, PubMed: 23664949

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x, PubMed: 31016684

Arsenault, J. S., & Buchsbaum, B. R. (2015). Distributed neural representations of phonological features during speech perception. *Journal of Neuroscience*, *35*(2), 634–642. https://doi.org/10.1523/JNEUROSCI.2454-14.2015, PubMed: 25589757

Avotec. (n.d.). *MRI audio products* [Apparatus]. https://www.avotecinc.com

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001, PubMed: 24403724

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, *14*(16), 2105–2109. https://doi.org/10.1097/00001756-200311140-00019, PubMed: 14600506

Belin, P., Zatorre, R. J., Hoge, R., Evans, A. C., & Pike, B. (1999). Event-related fMRI of the auditory cortex. *NeuroImage*, *10*(4), 417–429. https://doi.org/10.1006/nimg.1999.0480, PubMed: 10493900

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature, 403*(6767), 309–312. https://doi.org/10.1038/35002078, PubMed: 10659849

Bestelmeyer, P. E. G., Belin, P., & Grosbras, M.-H. (2011). Right temporal TMS impairs voice detection. *Current Biology, 21*(20), R838–R839. https://doi.org/10.1016/j.cub.2011.08.046, PubMed: 22032183

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: An fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience, 17*(9), 1353–1366. https://doi.org/10.1162/0898929054985473, PubMed: 16197689

Boersma, P., & Weenik, D. (2017). *Praat: Doing phonetics by computer.* https://www.fon.hum.uva.nl/praat/

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 27*(2), 113–120. https://doi.org/10.1109/TASSP.1979.1163209

Correia, J. M., Jansma, B., & Bonte, M. (2015). Decoding articulatory features from fMRI responses in dorsal speech regions. *Journal of Neuroscience, 35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015, PubMed: 26558773

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014, PubMed: 8812068

Crinnion, A. M., Malmskog, B., & Toscano, J. C. (2020). A graph-theoretic approach to identifying acoustic cues for speech sound categorization. *Psychonomic Bulletin and Review, 27*(6), 1104–1125. https://doi.org/10.3758/s13423-020-01748-1, PubMed: 32671571

Davis, M. H., Ford, M. A., Kherif, F., & Johnsrude, I. S. (2011). Does semantic context benefit speech understanding through "top–down" processes? Evidence from time-resolved sparse fMRI. *Journal of Cognitive Neuroscience, 23*(12), 3914–3932. https://doi.org/10.1162/jocn_a_00084, PubMed: 21745006

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., & Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage, 43*(1), 44–58. https://doi.org/10.1016/j.neuroimage.2008.06.037, PubMed: 18672070

Desai, R., Liebenthal, E., Waldron, E., & Binder, J. R. (2008). Left posterior temporal regions are sensitive to auditory categorization. *Journal of Cognitive Neuroscience, 20*(7), 1174–1188. https://doi.org/10.1162/jocn.2008.20081, PubMed: 18284339

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *Journal of the Acoustical Society of America, 144*(2), 1089–1099. https://doi.org/10.1121/1.5047672, PubMed: 30180678

Edmister, W. B., Talavage, T. M., Ledden, P. J., & Weisskoff, R. M. (1999). Improved auditory cortex imaging using clustered volume acquisitions. *Human Brain Mapping, 7*(2), 89–97. https://doi.org/10.1002/(SICI)1097-0193(1999)7:2<89::AID-HBM2>3.0.CO;2-N, PubMed: 9950066

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119*(4), 1950–1953. https://doi.org/10.1121/1.2178721, PubMed: 16642808

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex, 25*(12), 4772–4788. https://doi.org/10.1093/cercor/bhv136, PubMed: 26157026

Fant, G. (1968). Analysis and synthesis of speech processes. In B. Malmberg (Ed.), *Manual of phonetics* (pp. 173–277). North-Holland Publishing.

Feng, G., Yi, H. G., & Chandrasekaran, B. (2019). The role of the human auditory corticostriatal network in supervised speech learning. *Cerebral Cortex, 29*(10), 4077–4089. https://doi.org/10.1093/cercor/bhy289, PubMed: 30535138

Fischl, B. (2012). FreeSurfer. *NeuroImage, 62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021, PubMed: 22248573

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America, 106*(3), 1511–1522. https://doi.org/10.1121/1.427148, PubMed: 10489707

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science, 322*(5903), 970–973. https://doi.org/10.1126/science.1164318, PubMed: 18988858

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*(1), 110–125. https://doi.org/10.1037/0096-1523.6.1.110, PubMed: 6444985

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., & Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14*(3), 247–255. https://doi.org/10.1093/cercor/bhg124, PubMed: 14754865

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166, PubMed: 8926483

Greve, D. N. (2002). *Optseq* [Computer software]. https://surfer.nmr.mgh.harvard.edu/optseq/

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage, 180*(Pt A), 4–18. https://doi.org/10.1016/j.neuroimage.2017.08.005, PubMed: 28782682

Hebart, M. N., Görgen, K., & Haynes, J.-D. (2015). The Decoding Toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics, 8*, 88. https://doi.org/10.3389/fninf.2014.00088, PubMed: 25610393

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences, 4*(4), 131–138. https://doi.org/10.1016/S1364-6613(00)01463-7, PubMed: 10740277

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92*(1–2), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011, PubMed: 15037127

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113, PubMed: 17431404

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*(5), 3099–3111. https://doi.org/10.1121/1.411872, PubMed: 7759650

Jesse, A. (2021). Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(1), 184–194. https://doi.org/10.1037/xlm0000805, PubMed: 31855000

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin and Review*, *18*(5), 943–950. https://doi.org/10.3758/s13423-011-0129-2, PubMed: 21735330

Joanisse, M. F., Zevin, J. D., & McCandliss, B. D. (2007). Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. *Cerebral Cortex*, *17*(9), 2084–2093. https://doi.org/10.1093/cercor/bhl124, PubMed: 17138597

Johnson, K. A. (2008). Speaker normalization in speech perception. In D. Pisoni & R. Remez (Eds.), *The handbook of speech perception* (1st ed., pp. 363–389). https://doi.org/10.1002/9780470757024.ch15

Johnson, K. A., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384. https://doi.org/10.1006/jpho.1999.0100

Jones, A. B., Farrall, A. J., Belin, P., & Pernet, C. R. (2015). Hemispheric association and dissociation of voice and speech information processing in stroke. *Cortex*, *71*, 232–239. https://doi.org/10.1016/j.cortex.2015.07.004, PubMed: 26247409

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136. https://doi.org/10.2307/522229

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111–127. https://doi.org/10.1016/j.jml.2019.05.001

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 3933–3936). IEEE. https://doi.org/10.1109/ICASSP.2008.4518514

Kilian-Hütten, N., Valente, G., Vroomen, J., & Formisano, E. (2011). Auditory cortex encodes the perceptual interpretation of ambiguous sound. *Journal of Neuroscience*, *31*(5), 1715–1720. https://doi.org/10.1523/JNEUROSCI.4572-10.2011, PubMed: 21289180

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience*, *34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698, PubMed: 30619905

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001, PubMed: 16095588

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. https://doi.org/10.1073/pnas.0600244103, PubMed: 16537458

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008, PubMed: 19104670

Larson Davis. (2021). *Acoustic equipment* [Apparatus]. https://www.larsondavis.com

Lee, Y.-S., Turkeltaub, P., Granger, R., & Raizada, R. D. S. (2012). Categorical speech processing in Broca's area: An fMRI study using multivariate pattern-based analysis. *Journal of Neuroscience*, *32*(11), 3942–3948. https://doi.org/10.1523/JNEUROSCI.3814-11.2012, PubMed: 22423114

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279, PubMed: 4170865

Liebenthal, E., Desai, R., Ellingson, M. M., Ramachandran, B., Desai, A., & Binder, J. R. (2010). Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral Cortex*, *20*(12), 2958–2970. https://doi.org/10.1093/cercor/bhq045, PubMed: 20382643

Luthra, S. (2021). The role of the right hemisphere in processing phonetic variability between talkers. *Neurobiology of Language*, *2*(1), 138–151. https://doi.org/10.1162/nol_a_00028

Luthra, S., Correia, J. M., Kleinschmidt, D. F., Mesite, L. M., & Myers, E. B. (2020). Lexical information guides retuning of neural patterns in perceptual learning for speech. *Journal of Cognitive Neuroscience*, *32*(10), 2001–2012. https://doi.org/10.1162/jocn_a_01612, PubMed: 32662731

Luthra, S., Fox, N. P., & Blumstein, S. E. (2018). Speaker information affects false recognition of unstudied lexical-semantic associates. *Attention, Perception, and Psychophysics*, *80*(4), 894–912. https://doi.org/10.3758/s13414-018-1485-z, PubMed: 29473144

Luthra, S., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Neural substrates of subphonemic variation and lexical competition in spoken word recognition. *Language, Cognition and Neuroscience*, *34*(2), 141–169. https://doi.org/10.1080/23273798.2018.1531140, PubMed: 31106225

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception, and Psychophysics*, *83*(5), 2217–2228. https://doi.org/10.3758/s13414-021-02261-w, PubMed: 33754298

Luzzi, S., Coccia, M., Polonara, G., Reverberi, C., Ceravolo, G., Silvestrini, M., Fringuelli, F., Baldinelli, S., Provinciali, L., & Gainotti, G. (2018). Selective associative phonagnosia after right anterior temporal stroke. *Neuropsychologia*, *116*(Pt B), 154–161. https://doi.org/10.1016/j.neuropsychologia.2017.05.016, PubMed: 28506806

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*(Pt B), 179–193. https://doi.org/10.1016/j.neuropsychologia.2018.03.039, PubMed: 29614253

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7, PubMed: 22083660

Mattys, S. L., & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. *Perception and Psychophysics*, *70*(7), 1235–1242. https://doi.org/10.3758/PP.70.7.1235, PubMed: 18927006

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562. https://doi.org/10.1080/03640210802035357, PubMed: 21635345

McGettigan, C., & Scott, S. K. (2012). Cortical asymmetries in speech perception: What's wrong, what's right and what's left?

Trends in Cognitive Sciences, *16*(5), 269–276. https://doi.org/10.1016/j.tics.2012.04.006, PubMed: 22521208

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321. https://doi.org/10.1037/0278-7393.31.2.306, PubMed: 15755247

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219–246. https://doi.org/10.1037/a0022325, PubMed: 21417542

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*(1), 101–112. https://doi.org/10.1177/00238309060490010601, PubMed: 16922064

Mesgarani, N., Cheung, C., Johnson, K. A., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1011. https://doi.org/10.1126/science.1245994, PubMed: 24482117

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. https://doi.org/10.3758/BF03210878, PubMed: 2345691

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076, PubMed: 21924359

Myers, E. B. (2007). Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation. *Neuropsychologia*, *45*(7), 1463–1473. https://doi.org/10.1016/j.neuropsychologia.2006.11.005, PubMed: 17178420

Myers, E. B., Blumstein, S. E., Walsh, E., & Eliassen, J. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, *20*(7), 895–903. https://doi.org/10.1111/j.1467-9280.2009.02380.x, PubMed: 19515116.

Myers, E. B., & Mesite, L. M. (2014). Neural systems underlying perceptual adjustment to non-standard speech tokens. *Journal of Memory and Language*, *76*, 80–93. https://doi.org/10.1016/j.jml.2014.06.007, PubMed: 25092949

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44. https://doi.org/10.1016/j.bandl.2016.11.001, PubMed: 27898342

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, *109*(3), 1181–1196. https://doi.org/10.1121/1.1348009, PubMed: 11303932

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9, PubMed: 12948518

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x, PubMed: 21526138

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328. https://doi.org/10.1037/0278-7393.19.2.309, PubMed: 8454963

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Prolific Academic. (2022). *Prolific* [Research platform]. https://www.prolific.co/

R Core Team. (2022). *R: A language and environment for statistical computing* (Version 4.2.1) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raizada, R. D., & Poldrack, R. A. (2007). Selective amplification of stimulus differences during categorical processing of speech. *Neuron*, *56*(4), 726–740. https://doi.org/10.1016/j.neuron.2007.11.001, PubMed: 18031688

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331, PubMed: 19471271

Rogers, J. C., & Davis, M. H. (2017). Inferior frontal cortex contributions to the recognition of spoken words and their constituent speech sounds. *Journal of Cognitive Neuroscience*, *29*(5), 919–936. https://doi.org/10.1162/jocn_a_01096, PubMed: 28129061

Roland. (2022). *R-05 WAVE/MP3 recorder* [Apparatus]. https://www.roland.com/us/products/r-05/

Saad, Z. S., & Reynolds, R. C. (2012). SUMA. *NeuroImage*, *62*(2), 768–773. https://doi.org/10.1016/j.neuroimage.2011.09.016, PubMed: 21945692

Saltzman, D. I., & Myers, E. B. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, *28*(4), 1354–1364. https://doi.org/10.3758/s13423-021-01885-1, PubMed: 33742423

Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2014). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, *27*(2), 280–291. https://doi.org/10.1162/jocn_a_00707, PubMed: 25170793

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406. https://doi.org/10.1093/brain/123.12.2400, PubMed: 11099443

Siemens Medical Solutions. (2022). *3T Prisma scanner* [Apparatus]. https://www.siemens-healthineers.com

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments* (R package version 0.21–2) [Computer software]. https://CRAN.R-project.org/package=afex

Slavit, D. H. (1999). Phonosurgery in the elderly: A review. *Ear, Nose and Throat Journal*, *78*(7), 505–512. https://doi.org/10.1177/014556139907800713, PubMed: 10429327

Stevens, A. A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, *18*(2), 162–171. https://doi.org/10.1016/j.cogbrainres.2003.10.008, PubMed: 14736575

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain. 3-dimensional proportional system: An approach to cerebral imaging*. Thieme.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, and Psychophysics*, *77*(5), 1674–1684. https://doi.org/10.3758/s13414-015-0854-0, PubMed: 25824889

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America*, *128*(4), 2090–2099. https://doi.org/10.1121/1.3467771, PubMed: 20968380

Turkeltaub, P. E., & Branch Coslett, H. (2010). Localization of sublexical speech perception components. *Brain and Language*, *114*(1), 1–15. https://doi.org/10.1016/j.bandl.2010.03.008, PubMed: 20413149

Tzeng, C. Y., Nygaard, L. C., & Theodore, R. M. (2021). A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychonomic Bulletin & Review*, *28*(3), 1003–1014. https://doi.org/10.3758/s13423-020-01840-6, PubMed: 33443706

Van Lancker, D. R., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829–834. https://doi.org/10.1016/0028-3932(87)90120-5, PubMed: 3431677

von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48–55. https://doi.org/10.1016/S0926-6410(03)00079-X, PubMed: 12763191

von Kriegstein, K., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage*, *22*(2), 948–955. https://doi.org/10.1016/j.neuroimage.2004.02.020, PubMed: 15193626

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience*, *30*(2), 629–638. https://doi.org/10.1523/JNEUROSCI.2742-09.2010, PubMed: 20071527

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science*, *14*(2), 372–384. https://doi.org/10.1111/j.1467-7687.2010.00986.x, PubMed: 21479106

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2, PubMed: 28695541

Xie, X., & Myers, E. B. (2018). Left inferior frontal gyrus sensitivity to phonetic competition in receptive language processing: A comparison of clear and conversational speech. *Journal of Cognitive Neuroscience*, *30*(3), 267–280. https://doi.org/10.1162/jocn_a_01208, PubMed: 29160743

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096–1110. https://doi.org/10.1016/j.neuron.2019.04.023, PubMed: 31220442

Zatorre, R. J., Meyer, E., Gjedde, A., & Evans, A. C. (1996). PET studies of phonetic processing of speech: Review, replication, and reanalysis. *Cerebral Cortex*, *6*(1), 21–30. https://doi.org/10.1093/cercor/6.1.21, PubMed: 8670635