Brain & Language 165 (2017) 33–44

Contents lists available at ScienceDirect

Brain & Language

journal homepage: www.elsevier.com/locate/b&l

Voice-sensitive brain networks encode talker-specific phonetic detail

Emily B. Myers [a,b,c,d,*], Rachel M. Theodore [a,c,d]

[a] University of Connecticut, Department of Speech, Language, and Hearing Sciences, 850 Bolton Road, Unit 1085, Storrs, CT 06269-1085, United States
[b] University of Connecticut, Department of Psychological Sciences, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020, United States
[c] Haskins Laboratories, 300 George Street, Suite 900, New Haven, CT 06511, United States
[d] Connecticut Institute for the Brain and Cognitive Sciences, 337 Mansfield Road, Unit 1272, Storrs, CT 06269-1085, United States

ARTICLE INFO

Article history:
Received 5 April 2016
Revised 13 September 2016
Accepted 4 November 2016

Keywords:
Voice recognition
Phonetic processing
Superior temporal gyrus
VOT

ABSTRACT

The speech stream simultaneously carries information about talker identity and linguistic content, and the same acoustic property (e.g., voice-onset-time, or VOT) may be used for both purposes. Separable neural networks for processing talker identity and phonetic content have been identified, but it is unclear how a singular acoustic property is parsed by the neural system for talker identification versus phonetic processing. In the current study, listeners were exposed to two talkers with characteristically different VOTs. Subsequently, brain activation was measured using fMRI as listeners performed a phonetic categorization task on these stimuli. Right temporoparietal regions previously implicated in talker identification showed sensitivity to the match between VOT variant and talker, whereas left posterior temporal regions showed sensitivity to the typicality of phonetic exemplars, regardless of talker typicality. Taken together, these results suggest that neural systems for voice recognition capture talker-specific phonetic variation.

1. Introduction

Individual talkers differ in how they implement phonetic properties of speech (e.g., Allen, Miller, & DeSteno, 2003; Hillenbrand, Getty, Clark, & Wheeler, 1995; Newman, Clouse, & Burnham, 2001; Peterson & Barney, 1952; Theodore, Miller, & DeSteno, 2009). This kind of systematic talker-specific, within-category phonetic variation contributes to one's *idiolect*, that is, one's vocal identity. It has long been known that within-category phonetic variation is not discarded by the perceptual system; rather, it is used probabilistically to constrain and facilitate linguistic processing at both prelexical and lexical levels of representation (Andruski, Blumstein, & Burton, 1994; McMurray, Tanenhaus, Aslin, & Spivey, 2003; Myers, 2007; Pisoni & Tash, 1974; Utman, Blumstein, & Sullivan, 2001). Indeed, behavioral evidence suggests that listeners can perceptually track talker-specific phonetic variation and simultaneously use this information to identify *who* is doing the talking (e.g., Theodore & Miller, 2010) and to facilitate processing of *what* is being said (e.g., Nygaard, Sommers, & Pisoni, 1994). Of interest for the current study, the same acoustic cues (e.g., voice-onset-time values specifying stop consonants, formant patterns specifying vowels) are useful for both the *who* and *what* purposes (e.g., Theodore, Myers, & Lomibao, 2015). While much is known about the neural systems that underlie the processing of talker identity and those involved in processing the phonetic details of speech (e.g. Blumstein & Myers, 2014), it is unclear to what extent the neural systems that process talker information and phonetic information are dissociable or mutually interactive, particularly in the context of speech variants that contribute to a talker's idiolect. Below we review evidence from behavioral and neuroimaging paradigms that inform this question.

1.1. Interactive processing of phonetic and talker information

Behavioral examinations have revealed a tight link between the processing of phonetic and talker information (e.g., Theodore & Miller, 2010; Theodore et al., 2015). With respect to the processing of phonetic information, listeners receive comprehension benefits for familiar compared to unfamiliar talkers including heightened word recognition in degraded listening environments (e.g., Nygaard et al., 1994) and faster processing times (e.g., Clarke & Garrett, 2004). Research suggests that the processing benefits observed at higher levels of linguistic processing (e.g., word recognition) reflect adjustments that listeners make earlier in the perceptual stream (e.g., Nygaard & Pisoni, 1998). Listeners can learn a talker's characteristic VOT production for word-initial voiceless stops, indicating sensitivity to talker differences in individual pho-

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, University of Connecticut, 850 Bolton Rd, Unit 1085, Storrs, CT 06269-1085, United States.
E-mail address: emily.myers@uconn.edu (E.B. Myers).

http://dx.doi.org/10.1016/j.bandl.2016.11.001
0093-934X/

netic properties of speech (Allen & Miller, 2004; Theodore & Miller, 2010). Moreover, exposure to a talker's characteristic productions promotes a comprehensive reorganization of phonetic category structure such that behavioral judgments of phonetic category prototypicality reflect experience with individual talkers' characteristic productions of those phonetic categories (Theodore et al., 2015). With respect to the processing of talker information, research has shown that voice recognition is heightened in the native compared to a non-native language (Goggin, Thompson, Strube, & Simental, 1991; Xie & Myers, 2015). The language familiarity benefit for voice recognition has been linked to experience and expertise with phonetic variation associated with linguistic sound structure (Johnson, Westrek, Nazzi, & Cutler, 2011; Orena, Theodore, & Polka, 2015). As argued in Perrachione and Wong (2007), the reliance of talker identification processes on phonetic qualities of the input points to a shared neural substrate for perception of talker identity and phonetic characteristics. Collectively, these behavioral findings demonstrate that the processing of phonetic and talker information are fundamentally linked with respect to spoken language processing, and further suggest that their linkage emerges at a sublexical level of representation.

1.2. Neural systems for processing phonetic variation and talker information

The neural systems that extract phonetic content from the speech signal and those that extract information about vocal identity are partially overlapping, but can be argued to recruit different circuits (Blumstein & Myers, 2014; Van Lancker, Kreiman, & Cummings, 1989). For the purposes of phonetic processing, within-category phonetic variability is processed in the bilateral posterior superior temporal gyrus (STG) (see Blumstein & Myers, 2014 for review; Chang et al., 2010; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Myers, 2007). Of interest, regions in the bilateral superior temporal lobes show tuning to the best exemplars of one's native language phonetic category, with greater activation observed for phonetic tokens that are less typical as a member of the category (e.g., a /t/ with an extremely long VOT) compared to more standard productions (Myers, 2007). These core phonetic processing regions are permeable to at least some top-down influences. For example, these regions show differences in sensitivity to tokens along a phonetic continuum when the phonetic category boundary has been shifted by embedding a token in a biasing lexical or sentential context (Gow, Segawa, Ahlfors, & Lin, 2008; Guediche, Salvata, & Blumstein, 2013; Myers & Blumstein, 2008). This said, it is unclear whether every source of information—and in particular, whether a given token is typical of a *talker's* voice—modulates the tuning of the STG to native language typicality.

While phonetic processing is thought to be bilateral, with some preference for leftwards laterality, processing of vocal identity has largely been attributed to right hemisphere regions (e.g. Van Lancker et al., 1989; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). A classic study by Van Lancker et al. (1989) tested a group of individuals with left and right hemisphere lesions on identification of familiar voices as well as voice discrimination. Voice discrimination was impaired in individuals with both left and right hemisphere temporal lesions, but identification of familiar voices was impaired in individuals with right inferior parietal lesions (see also Van Lancker, Cummings, Kreiman, & Dobkin, 1988). Imaging studies have further corroborated the separation between regions that are sensitive to the acoustics of the voice—and thus could be used for discriminating between talkers—and those responsible for mapping voice acoustics to an individual identity which can be used for talker identification (von Kriegstein et al., 2003). In particular, while voice acoustics may be processed in bilateral temporal regions (specifically the superior temporal sulcus or STS), imaging studies have sited vocal identity processing (or access to familiar voices) in the anterior right temporal lobe rather than the right posterior region implicated in lesion studies (Andics, McQueen, & Petersson, 2013; Andics et al., 2010; Belin & Zatorre, 2003; Campanella & Belin, 2007).

Other evidence corroborates the role of either right anterior temporal or posterior temporoparietal regions for processing vocal identity (Belin & Zatorre, 2003; Stevens, 2004; von Kriegstein et al., 2003). For instance, von Kriegstein et al. (2003) showed that shifts in attention to vocal identity resulted in shifts in activation to right STS (see also, Belin & Zatorre, 2003), and revealed a gradient of processing such that anterior regions did not differentiate between familiar and unfamiliar voices, whereas posterior regions responded more to unfamiliar than familiar voices (Kriegstein & Giraud, 2004). A role for the right hemisphere in processing vocal identity is also supported by evidence that right frontal (middle frontal gyrus, MFG) and right inferior parietal regions (angular gyrus, AG) respond to short-term memory for talker identity (Stevens, 2004). Other studies have implicated bilateral temporal structures in processing changes in vocal identity when the linguistic message was held constant (Salvata, Blumstein, & Myers, 2012; Wong, Nusbaum, & Small, 2004). Of note, these latter studies cannot determine whether regions that respond to changes in vocal identity are those responsible for processing that identity itself, or whether they instead respond to other characteristics of the stimuli (e.g., differences in sensitivity to low-level acoustic properties that happen to differ across talkers). Notably, regions that respond to voice acoustics are more likely to be shared with the linguistic system, simply because the same properties of the acoustic signal can carry information about talker identity as well as linguistic content. The co-dependence of shared acoustic cues for phonetic processing and talker identity is highlighted in a study by von Kriegstein, Smith, Patterson, Kiebel, and Griffiths (2010). In this study, the perception of vocal tract length was manipulated by shifting the formant structure of utterances—crucially, this manipulation results in a change in the percept of talker identity that is signaled by an acoustic cue (formant structure) that is also used for vowel identity. In this study, the right STS/STG response to perceived vocal tract length (or talker identity) was amplified when participants were engaged in a task that focused attention at the phonetic level. Taken together, these studies suggest that neural systems arrayed along the right temporal lobe process talker-level information.

Andics et al. (2010) proposed that not only are the processing stages that process voice acoustics and voice identity separable, but that there are two sets of neural coding spaces, a 'voice-acoustics' space and 'voice-identity' space, each of which codes prototypical members of that space more sparsely then items distant from the prototype. This research group showed that the same core temporal lobe regions found to be sensitive to phonetic processing in other studies (see Myers, 2007; Myers, Blumstein, Walsh, & Eliassen, 2009) were sensitive to the internal structure of a learned voice identity space, showing less activation for stimuli that were more prototypical of a learned voice and greater activation for stimuli that were less typical of the talker's voice. Notably, in this study, the stimulus space was defined as a morph between two talker's voices, and therefore the continuum was likely to vary in voice-diagnostic features such as pitch contour and timbre, but potentially also in phonetic variation between the two talkers, although these details were not specified in the report. Nonetheless, a separate set of regions in anterior temporal areas was found to be sensitive to talker identity when controlling for the acoustic distinctions along the continuum, and only these anterior temporal regions correlated with identification performance. The authors take this finding as evidence that regions that

are sensitive to vocal identity are separable from those sensitive to vocal acoustics.

Thus far, evidence reviewed above suggests that acoustic variability within the phonetic category is processed within regions of the brain associated with phonetic processing (especially the bilateral STG), and that a network of right-lateralized neural regions are sensitive to talker identity. What is unclear is whether the neural systems responsible for processing phonetic variability in order to map these sounds to meaning, and those processing the same cues to map to talker identity, are separable or overlapping. That is, many acoustic cues combine to contribute to talker recognition, including cues that can be isolated from the phonetic properties of speech, such as fundamental frequency (F_0), timbre, jitter, and shimmer (see Creel & Bregman, 2011 for review; Gelfer, 1988; Van Lancker et al., 1989). Other properties, such as VOT and formant patterns, are used for both talker identification and phonetic processing. Previous studies have investigated the overlap between phonetic and talker identity cues by manipulating factors such as formant structure that have a predictable relationship to a potential anatomical configuration (i.e., vocal tract length), and thus have a fairly deterministic relationship to talker identity (see von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007; von Kriegstein et al., 2010). However, to our knowledge, no studies have investigated acoustic cues that are idiosyncratic to the talker such as VOT, and that may be adopted by talkers irrespective of their vocal tract anatomy.

A study from our group provides some hints at the neural processes that link acoustic variation to a talker's vocal identity (Myers & Mesite, 2014). This investigation of an effect termed "lexically-guided perceptual learning" (Kraljic & Samuel, 2007; Norris, McQueen, & Cutler, 2003) used lexical information to bias the perception of an ambiguous phoneme (in this case, a blend of /s/ and /ʃ/) towards either an /s/ interpretation or an /ʃ/ interpretation by embedding this ambiguous phoneme in the context of words that were consistent with only one of these phonemes (e.g.,. in place of the /s/ in *episode* or the /ʃ/ in *flourishing*). Sensitivity to this shift in category boundary was seen in regions that had been previously implicated in access to voice recognition systems (right MTG and right MFG). Although listeners were only exposed to one talker's voice in this study, it may be that the idiosyncratic, ambiguous token that appeared in this stimulus set became a part of the vocal identity representation for that talker, which is consistent with other findings suggesting that the nature of the perceptual adaptation in this paradigm is often talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2007). Further, in this study, the driving mechanism behind changes in sensitivity to the phonetic continuum was a top-down influence from lexical status: that is, the identity of the ambiguous phoneme had to be resolved in order to access the word in which it was embedded (i.e., to understand that one is hearing the real word *episode* and not the non-word *epishode*). It is unclear whether the shifts in talker-specific phonetic retuning seen in core phonetic processing regions are limited to those cases in which top-down linguistic information exerts an influence on phonetic category processing.

The goal of the current study is to answer two key questions: (1) do neural regions associated with processing phonetic typicality show talker-specific modulation as a function of experience with a talker's characteristic productions, and (2) do the networks in right anterior temporal lobe and right temporoparietal areas that respond to vocal identity show modulation that reflects talker-specific phonetic variation? In order to answer these questions, listeners were exposed to two talkers whose VOTs were modified such that one talker always employed longer VOTs and the other always used shorter VOTs for the voiceless stop /k/. Notably, these differences in VOT are all clear tokens of /k/, and there is no sense in which associating this variant with a particular talker will resolve any ambiguities in the input (see Kraljic & Samuel, 2007; Norris et al., 2003). Moreover, the two talkers had perceptually distinct voices that differed in standard vocal cues (e.g., F_0, timbre) in addition to their difference in the phonetic property of VOT. Participants were trained to recognize these talkers, and neural sensitivity to the typicality of a VOT variant as characteristic of a talker's voice was measured using fMRI as participants performed a phonetic categorization task.

2. Material and methods

2.1. Participants

Seventeen right-handed, monolingual native speakers of English participated in the MRI study. Participants were recruited from the Brown University community, gave informed consent according to the regulations of the Brown Institutional Review Board, and were compensated at a rate of $20 per hour. Participants reported normal hearing and no history of any neurological or language disorder, and were screened for the presence of bodily ferromagnetic materials. Participants were randomly assigned to one of two training groups, either the S-SHORT/J-LONG training group or the S-LONG/J-SHORT training group. The data from two participants were excluded; one participant showed near-chance performance on the in-scanner task, and a second participant showed movement in excess of 6 mm during the functional scans. Data from the remaining fifteen participants (8 female, mean age = 27 - years, SD = 7 years, range = 18–42 years) are reported below, eight in the S-SHORT/J-LONG training group and seven in the S-LONG-J-SHORT training group.

2.2. Stimuli

The stimuli were drawn from those used in Theodore and Miller (2010), to which the reader is referred for comprehensive details of stimulus construction. In brief, the stimuli for the current experiment consisted of tokens from two synthesized VOT continua that ranged from *gain* to *cane*, one for each of two female speakers, referred to as Sheila and Joanne. The two talkers were monolingual speakers of American English with perceptually distinct voices. The continua were created using an LPC-based speech synthesizer and used a naturally-produced token of *gain* as the endpoint for each continuum. The selected *gain* tokens were equated for word duration (568 ms) and root-mean-square amplitude (RMS). Successive steps on each continuum were generated by editing parameters of the LPC analysis of the original token in order to change the periodic source to a noise source for successive pitch periods, thus increasing VOT in very small steps on each successive token. This procedure thus yielded, for each talker, numerous tokens that unambiguously cued the words *gain* and *cane*, but with different VOTs in each case.

Drawing from these continua, sets of tokens were selected for use during a talker training task, a talker typicality judgment task, and a phonetic categorization task. The final stimulus set is schematized in Fig. 1. For the talker training task, we selected the following from each talker: one *gain* token, two *cane* tokens with short VOTs that were two steps apart on the continuum, and two *cane* tokens with relatively longer VOTs that were also two steps apart. VOTs were matched between the two talkers for each type of token (*gain*, Short VOT *cane*, Long VOT *cane*). The selected tokens were organized into two sets, one for each training group. The S-SHORT/J-LONG training group used Sheila's Short VOT *cane* tokens, Joanne's Long VOT *cane* tokens, and the *gain* tokens from both speakers. The S-LONG/J-SHORT training group used Sheila's Long VOT *cane* tokens, Joanne's Short VOT *cane* tokens,

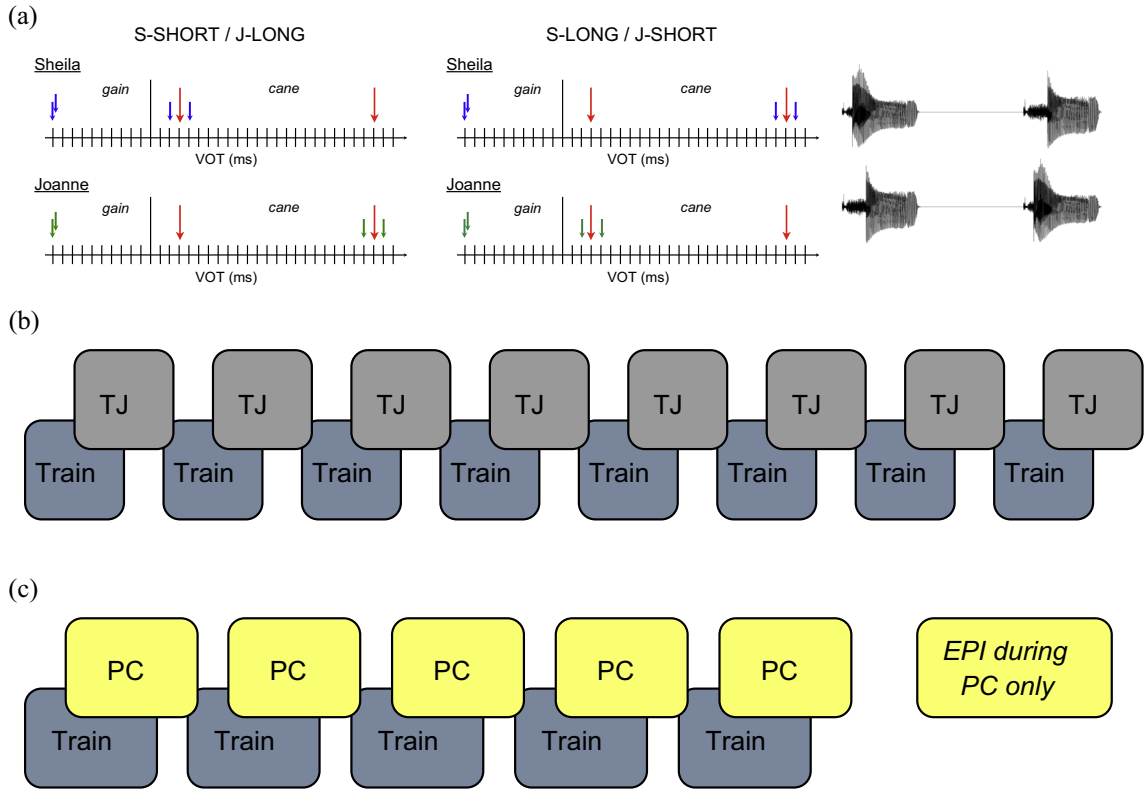

Fig. 1. Schematic of experimental design. Panel (a) depicts tokens along the VOT continuum used as stimuli in the Talker Training (Train) task and as test stimuli in the Talker Typicality Judgment (TJ) task and Phonetic Categorization (PC) tasks. Blue arrows indicate tokens selected for Sheila's voice; green arrows indicate tokens selected for Joanne's voice. Note that the test stimuli, shown in red, were the same for the two training groups; representative waveforms (indicating amplitude as a function of time) for the TJ task are shown at right in panel (a). Panel (b) shows the alternating pattern of tasks used during the pre-scanner behavioral session and panel (c) shows the alternating pattern of tasks used during the scanning session.

and the *gain* tokens for both speakers. In each training set, we duplicated the *gain* token for each speaker to equate the number of *gain* and *cane* items. We also created two amplitude variants for each selected token, corresponding to the RMS amplitude of the Short VOT and Long VOT variants, respectively, in order to remove a potential amplitude-based confound (Theodore & Miller, 2010). With this design, each talker training set consisted of 16 tokens.

For the talker typicality judgment task, we created one set of stimuli that was used for both training groups, consisting of a Short VOT and a Long VOT variant of *cane* for each talker. Recall that for the talker training tokens, the selected Short VOT and Long VOT variants were each two steps apart on the continuum; the intermediate tokens were used for the talker typicality judgment. Two amplitude variants of these tokens were created corresponding to mean RMS amplitude of the selected Short VOT and Long VOT variants of *cane* used during training. We then created pairs of test stimuli by concatenating a Short VOT and a Long VOT variant, separated by 750 ms of silence. This procedure yielded four test pairs for each talker, half that began with the Short VOT token and half that began with the Long VOT token, with amplitude held constant for a given pair. Thus, there were eight talker typicality judgment stimuli in total, four for Sheila's voice and four for Joanne's voice.

For the phonetic categorization task, we again created one set of stimuli that was used for both training groups, consisting of the eight *gain* tokens used in the talker training phase and the eight *cane* tokens used during the talker typicality judgment task, for a total of 16 tokens for use during phonetic categorization. With this design, both test tasks (talker typicality, phonetic categorization) use the same *cane* stimuli, which are physically distinct from those presented during the talker training task.

2.3. Behavioral procedure

After providing informed consent and undergoing safety screening for MR compatibility, participants completed the talker training task and the talker typicality judgment task outside the scanner in order to provide the participants with systematic exposure to each talker's characteristic VOT productions, modeled after the procedures used in Theodore and Miller (2010) and Theodore et al. (2015). Both groups of listeners heard the two talkers produce the words *gain* and *cane*; however, we manipulated exposure such that one group of listeners heard Sheila produce /k/ with short VOTs and Joanne produce /k/ with relatively longer VOTs (S-SHORT/J-LONG training group) and the other group of listeners heard the opposite pattern of VOT exposure (S-LONG/J-SHORT training group). In order to provide systematic exposure and confirm that listeners were indeed tracking the talker differences in characteristic VOT production, all participants completed the talker training task and the talker typical judgment task, described in detail below. The overall procedure required listeners to alternate between the two tasks and is illustrated in the experimental timeline shown in Fig. 1.

2.3.1. Talker training

Participants were trained on a set of voiced (*gain*) and voiceless (*cane*) stimuli in both talkers' voices as appropriate for their experimental group (S-SHORT/J-LONG vs. S-LONG/J-SHORT) such that each talker had a characteristic VOT associated with her voice. During the talker training task, participants performed a four-alternative forced-choice (4AFC) task in which they were presented with one word at a time and asked to simultaneously decide the talker identity and the word itself and press a corresponding

button (i.e., Sheila/Gain, Sheila/Cane, Joanne/Gain, Joanne/Cane). In total, listeners completed 8 blocks of the talker training task in the behavioral session, with each block consisting of one randomization of the 16 training stimuli.

2.3.2. Talker typicality judgment

In addition to the talker training blocks, participants performed a talker typicality judgment task on tokens drawn from each talker's voice. On each trial, listeners heard two VOT variants of the word *cane* produced by one of the two talkers, one was a Short VOT variant and the other was a Long VOT variant, with the order of the two variants randomized across trials. For each trial, listeners were asked whether the first or second variant in the pair sounded more characteristic of the talker's voice. Each block of the talker typicality judgment consisted of eight trials, and listeners completed four blocks for each talker's voice throughout the behavioral session. Participants were queried about each talker's voice on alternating blocks, with the order counterbalanced across subjects.

In total, the behavioral training outside the scanner consisted first of a familiarization block in which listeners heard a single word randomly drawn from the set of 16 training stimuli and saw the talker's name ("Joanne" or "Sheila") displayed on the screen. No response was required from the subject. Following familiarization, participants practiced the talker typicality judgment task on each talker's voice. Finally, training consisted of alternating blocks of the talker training and talker typicality judgment tasks, to the completion of eight cycles of these two tasks.

2.3.3. Phonetic categorization

Following the behavioral training, participants entered the scanner, performed one short block of the talker training task while an anatomical scan (MPRAGE) was acquired, and then performed a phonetic categorization task during functional (EPI) scans. Importantly, during scanning itself, participants were not asked to make any judgment about talker identity or talker typicality, but instead simply categorized each stimulus as either *cane* or *gain*. Participants listened to stimuli over noise-attenuating MR-compatible headphones (Avotech) and indicated their response with an MR-compatible button box placed under the right hand. Participants heard three randomizations of the 16 tokens selected for use during the phonetic categorization task, presented one at a time; thus, each run consisted of 48 trials. Half of the trials were in each talker's voice, and of these, half were voiced variants (*gain*) and half were voiceless variants (*cane*). Crucially, of the voiceless stimuli, half (n = 6 per run) were at a VOT value that was typical for that talker's voice and half were atypical. For instance, for participants who had been trained with Short VOT variants for Sheila (the S-SHORT/J-LONG training group), Sheila-Short trials were typical whereas Sheila-Long trails were atypical, with the opposite mapping for Joanne's voice. There were five functional runs of the phonetic categorization task, with trials jittered at multiples of the TR, for a total of 30 trials in each of the critical cells (Sheila-Short, Sheila-Long, Joanne-Short, Joanne-Long).

During pauses between each of five functional runs, participants performed a refresher mini-block of the talker training task in order to minimize the possibility that exposure to non-standard variants for each talker during the phonetic categorization task might erode the talker-to-variant mapping, as shown in Fig. 1.

2.4. MRI procedure and analysis

MRI data was acquired using a Siemens 3T Tim Trio scanner at Brown University, using a 32-channel head coil. Anatomical MPRAGE scans were acquired for image co-registration (1 mm^3 voxels, 160 slices, TR = 1.9 s TE = 2.98 ms). Functional scans consisted of five echo-planar (EPI) runs of 123 (first ten subjects) or 120 volumes (final five subjects) each. EPI data were acquired using a clustered acquisition sequence which allowed for the presentation of audio stimuli during the relative silence between functional scans (Edmister, Talavage, Ledden, & Weisskoff, 1999) and each TR consisted of a 2000 ms scan followed by 800 ms of silence, during which the stimuli were presented. Phonetic categorization trials were presented with SOAs spaced in multiples of the TR multiples ranging from 2.8 s to 11.2 s, with an average spacing of 7 s. During each TR, 27 slices were acquired in oblique orientation with slices oriented parallel to the sylvian fissure, with an in-plane resolution of 2 mm^2. For one subject (14SS) slice thickness was 2 mm and for the remaining participants slice thickness was 3 mm (TR = 2.8 s, TE = 32 ms). The data were processed using AFNI (Cox, 1996). Functional datasets were reconstructed, resampled at 2 mm^3, and aligned to the anatomical images. Functional data were motion corrected using the fourth volume as a reference volume, spatially smoothed using a 4 mm Gaussian kernel, and converted to percent signal change units. The first two volumes of every run were discarded to avoid saturation artifacts, outlier volumes were censored from further analysis, and data were masked to include only those voxels in which at least 11 out of 15 participants had measurable data. Anatomical images were skull-stripped and transformed to Talairach space (Talairach & Tournoux, 1988).

2.4.1. Subject-level analysis

First-level analysis was performed using AFNI's 3dDeconvolve program. A stereotypic hemodynamic function was convolved with stimulus start times for each condition of interest and by-subject voxel-wise beta coefficients were estimated for each condition. Separate regressors were included for Short Voiceless, Long Voiceless, and Voiced stimuli separately for Joanne and Sheila's voices, for a total of six regressors.

2.4.2. Group-level analysis

Group-level analysis was performed in two ways. First, beta coefficients from the voiceless trials only were entered into a 2×2 ANOVA in order to explicitly isolate the effects of the VOT variant, with Talker (Joanne vs. Sheila) and VOT (Long vs. Short) as the factors. Second, in order to isolate effects of the typicality of the VOT variant for each talker's voice, stimulus conditions were recoded, with talker as one factor (Joanne vs. Sheila) and Typicality as the other factor (Typical vs. Atypical). Note that for this analysis, both Long and Short VOT variants were included in each of the Typical and Atypical codes because across participants, the VOT variant that was typical for that talker was counterbalanced. This second analysis allows us to control for differences that might arise due to typicality of Long vs. Short VOTs more generally. In both cases, the talker voice (Joanne vs. Sheila) is of no theoretical interest, in the sense that any differences between the two talkers could be triggered by low-level differences in the acoustic properties of these two voices (e.g., pitch variation, timbre). However, because this factor was included in the analysis, the contrast between these two voices is reported in Table 1. Three directional *t*-tests are reported, with group-level results corrected for multiple comparisons using two voxel-level threshold and cluster size combinations, each of which achieved a cluster-corrected threshold of $p < 0.05$ according to Monte Carlo simulations performed using AFNI's 3dClustSim tool. Group results were displayed on a surface reconstruction of a standard brain using SUMA (Saad & Reynolds, 2012).

2.4.3. gPPI analysis

To examine task-related changes in connectivity to a region of interest (ROI) that was identified to be responsive to typicality of

Table 1
Clusters from three planned *t*-tests. All clusters significant at a cluster-corrected p < 0.05. Coordinates are given in Talariach space.

Cluster size in voxels	Peak x	Peak y	Peak z	Region	*t*-statistic at peak
			Typical > Atypical		
323	43	−55	25	R STG, RMTG, RSMG	4.46
192	1	−45	27	L post Cingulate, L Cingulate	4.39
			Long > Short		
359	−3	−18	13	R Cuneus	5.56
268	−57	36	−7	R MTG, R STG	3.48
227	39	26	18	L MTG, L STG	4.78
222	3	−34	21	Ant. Cingulate	5.21
			Sheila > Joanne		
254	25	−47	32	R Inferior Parietal Lobule	6.57
161	11	−55	35	R Precuneus	3.65
142	−45	6	−10	L STG, L IFG	4.77
136	−41	41	−1	L MFG	5.00
135	−50	−5	−14	L MTG, LSTG	5.56

phonetic tokens (Atypical vs. Typical), a generalized psychophysiological interaction (gPPI) analysis was performed. PPI analyses identify regions in which connectivity to a seed region is modulated by the task—in this case, the "typicality" of the stimulus a participant is hearing (McLaren, Ries, Xu, & Johnson, 2012). First, the by-subject time course within a seed ROI identified in the Group Level analysis (the right MTG area identified in the Typical vs. Atypical contrast) was extracted and detrended, removing Legendre polynomials up to 5th-order trends, and then deconvolved using a stereotypic hemodynamic response. Second, a regressor expressing the interaction between each condition and the seed region was created by multiplying the condition regressors for each condition by the seed time course. Finally, the seed time course, original condition regressors, and the condition by seed interaction regressors were all entered into a by-subject regression analysis, which allowed us to identify regions showing a significant relationship to the interaction of the seed region and the stimulus condition, over and beyond any variance attributable to general correlation to the seed time course or condition-related activation. Group analysis entailed performing a *t*-test on the PPI results for Typical compared to Atypical trials. The statistical map was thresholded using the same criteria as the functional analysis reported above.

3. Results

3.1. Behavioral results

3.1.1. Pre-scan training performance

Results during the pre-scan training protocol were analyzed separately for the talker training and talker typicality judgment tasks. For the talker training task, mean percent correct talker identification was calculated for each participant separately for the two voices by collapsing over the eight talker training blocks. In the same fashion, mean percent correct phonetic decision during this task was calculated for each participant separately for the two voices. In all cases, mean performance for each training group was near ceiling and is shown in Fig. 2a. These results confirm that the listeners learned the talkers' voices and that the VOT variants were perceived as intended (e.g., the Short VOT variants were perceived as /k/ and not as /g/) during the talker training task.

Recall that for the talker typicality judgment task, listeners were directed to select either the Short VOT or Long VOT variant on each trial. For each participant, we measured mean percent VOT-responses separately for each talker by collapsing over the four test blocks for each talker. Mean performance for the two training groups is shown in Fig. 2b. As expected, listeners demonstrated sensitivity to each talker's characteristic productions such that which variant was selected at test was in line with their exposure during training. Specifically, listeners who heard Sheila produce long VOTs during the talker training selected more Long VOT variants during the talker typicality judgment task compared to listeners who heard Sheila produce short VOTs during training; similarly, responses to Joanne's voice show exposure-dependent decisions.

To examine this pattern statistically, mean percent Long VOT responses was submitted to ANOVA with the between-subjects factor of training group (S-SHORT/J-LONG vs. S-LONG/J-SHORT) and the within-subjects factor of talker (Sheila vs. Joanne). The ANOVA showed no main effect of training group [$F(1,13) = 2.79$, $p = 0.119$, $\eta_p^2 = 0.177$], no main effect of talker [$F(1,13) = 1.29$, $p = 0.276$, $\eta_p^2 = 0.090$], but a robust interaction between training group and talker [$F(1,13) = 92.45$, $p < 0.001$, $\eta_p^2 = 0.877$]. The nature of the interaction was confirmed statistically with two sets of planned comparisons. Independent *t*-tests confirmed that for Sheila's voice, there were fewer Long VOT responses for listeners in the S-SHORT/J-LONG (mean = 16.21%, SD = 18.33) compared to the S-LONG/J-SHORT training group (mean = 92.09%, SD = 16.10) [$t(13) = -8.46$, $p < 0.001$, $d = -4.69$], and that for Joanne's voice, there were more Long VOT responses for listeners in the S-SHORT/J-LONG (mean = 74.45%, SD = 20.68) compared to the S-LONG/J-SHORT training group (mean = 9.37%, SD = 9.19) [$t(13) = 7.66$, $p < 0.001$, $d = 4.25$]. Paired t-tests confirmed that listeners in the S-SHORT/J-LONG training group had fewer Long VOT responses for Sheila compared to Joanne [$t(7) = -5.28$, $p < 0.001$, $d = -2.98$], and that listeners in the S-LONG/J-SHORT training group had more Long VOT responses for Sheila compared to Joanne [$t(6) = 8.91$, $p < 0.001$, $d = -6.31$]. Consistent with previous findings (Theodore & Miller, 2010; Theodore et al., 2015), the current results indicate that listeners are sensitive to talker-specific phonetic detail such that they can learn a talker's characteristic VOT productions.

3.1.2. In-scanner performance

Recall that participants performed two tasks in the scanner including (1) the talker training task which was identical to that performed in the pre-scan training period and completed as a refresher between scans, and (2) a phonetic categorization task during which functional activation was measured. In order to confirm that the listeners retained learning from the pre-scan training protocol during the MRI procedures, we analyzed behavioral performance during the talker training task that was completed in the scanner. Specifically, we calculated mean percent correct talker decision and mean percent correct phonetic decisions for each participant separately for each talker's voice, as described above. Mean performance for both decisions was near ceiling for both

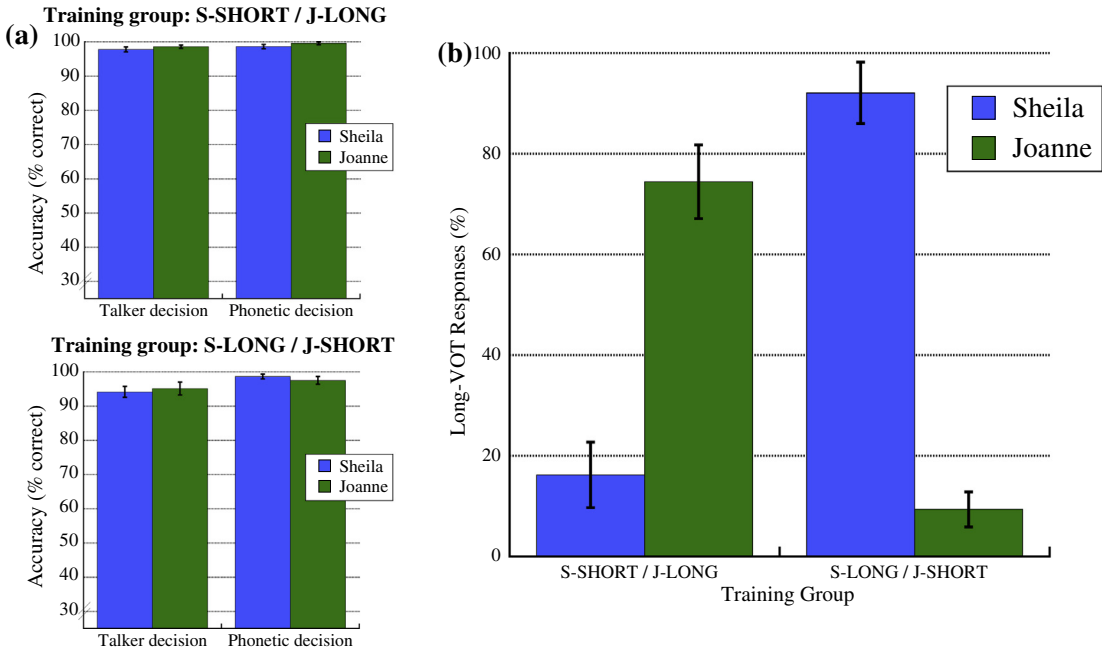

Fig. 2. Behavioral performance during the pre-scanner training session. For each training group, panel (a) shows mean percent correct talker and phonetic decisions for each talker from the talker identification task used in the pre-scan training protocol. Panel (b) shows mean percent Long VOT responses to each talker for each training group during the talker typicality judgment task in the pre-scan training protocol. Error bars in both panels indicate standard error of the mean.

talkers and for both training groups (mean > 95% in all cases), indicating that the learning that occurred outside of the scanner was retained during the scan.

Behavioral performance during the phonetic categorization task was measured in terms of accuracy (% correct) and response time. Consider first accuracy. A response was considered correct if the phonetic categorization decision was consistent with the intended speech sound; that is, a response was correct if the voiced tokens were identified as /g/ and the Short VOT and Long VOT variants were identified as /k/. For each participant, we calculated mean accuracy separately for each talker and for each stimulus type (voiced, Short VOT, Long VOT) by collapsing across the two amplitude variants of each stimulus type. (We note that four participants were excluded from the accuracy analysis due to a programming error with the button box response options.) As expected based on the behavioral data obtained in the talker training task, performance in all cases was near ceiling (mean > 94% in all cases), indicating that listeners perceived the test tokens as intended. Mean accuracy was submitted to ANOVA with the between-subjects factor of training group and the within-subjects factors of talker (Sheila vs. Joanne) and stimulus type (voiced vs. Short VOT vs. Long VOT). No main effects or interactions were statistically reliable ($p > 0.170$ in all cases).

Having confirmed in the accuracy analysis that all listeners perceived both talkers' Short VOT and Long VOT variants as /k/, the reaction time analyses were conducted in order to provide a more fine-grained measure of processing between these two variants. Recall that the Short VOT variants pattern more closely to VOT values produced in natural speech, with the values of the Long VOT variants being less prototypical exemplars. This raises the possibility that processing the Long VOT variants may be de facto more difficult than processing the Short VOT variants, which has implications for interpretation of the MRI analyses (as we elaborate further, below). As described below, the MRI analyses compared listeners' performance for each talker's Short VOT and Long VOT variants as a function of previous exposure to their voices; accordingly, we analyzed reaction time to the Short VOT and Long VOT variants in order to determine if this dependent measure would reveal differences in processing that were not captured in terms of response accuracy (which was near ceiling for both VOT variants and for both speakers). Mean reaction time was calculated for each participant separately for each talker and each voiceless VOT variant; mean performance for each training group is shown in Fig. 3.

We submitted mean reaction time to ANOVA with the between-subjects factor of training group (S-SHORT/J-LONG vs. S-LONG/J-SHORT) and the within-subjects factors of talker (Sheila vs. Joanne) and VOT variant (Short VOT vs. Long VOT). None of the main effects were reliable, indicating that mean reaction time was equivalent between the two training groups [$F(1,13) = 0.13$, $p = 0.912$, $\eta_p^2 = 0.001$], between the two talkers [$F(1,13) = 0.132$, $p = 0.722$, $\eta_p^2 = 0.010$], and between the two VOT variants [$F(1,13) = 0.489$, $p = 0.497$, $\eta_p^2 = 0.036$]. Moreover, none of the two-way interactions were reliable ($p > 0.250$ in all cases). Strikingly, the three-way interaction between training group, talker, and VOT variant was marginally reliable [$F(1,13) = 4.58$, $p = 0.052$, $\eta_p^2 = 0.260$]. As shown in Fig. 3, this interaction reflects a numerical trend for response time to decrease for variants that are typical of previous exposure compared to variants that are atypical. Post-hoc planned comparisons were conducted in order to confirm the nature of the three-way interaction in terms of eight paired *t*-tests that for each training group examined performance between the two talkers for each VOT variant and between each VOT variant for each talker. However, none of the planned comparisons reach statistical significance when applying the Bonferonni correction for family-wise error rate ($\alpha = 0.006$). Accordingly, the three-way interaction in the omnibus ANOVA should be interpreted with caution.

3.2. MRI results

3.2.1. Effects of phonetic category structure

This analysis was designed to select brain regions that respond to within-category phonetic variation, regardless of whether this variation was typical or atypical for a given talker. Previous work (Myers, 2007) showed that regions in the bilateral STG, extending

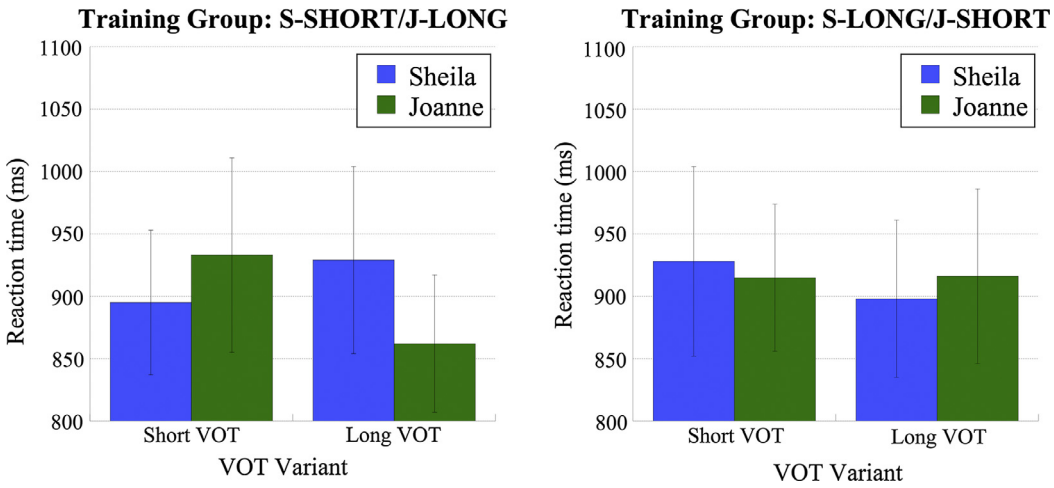

Fig. 3. Behavioral performance during the phonetic categorization task performed in the scanner as measured by reaction time (ms) to the Short VOT and Long VOT tokens. The left panel shows performance for the S-SHORT/J-LONG training group and the right panel shows performance for the S-LONG/J-SHORT training group. Error bars in both panels indicate standard error of the mean.

to middle temporal gyrus (MTG), were sensitive to the 'goodness of fit' of tokens to their phonetic category. A contrast in the current study of Long VOT vs. Short VOT variants yielded four clusters, two of which were centered in regions implicated in fine-grained acoustic-phonetic sensitivity. Specifically, more activation for Long VOT than Short VOT variants was seen in bilateral clusters spanning the MTG and STG, as well as within a right subcortical cluster and another in the anterior cingulate (see Table 1, Fig. 4).

3.2.2. Effects of talker typicality

In this analysis, the Long and Short VOT variants were recoded according to whether they were typical or atypical of a talker's speech. A main effect of typicality (Atypical vs. Typical) yielded two significant clusters, one in the right temporoparietal junction, and the second in the posterior cingulate (Fig. 5, Table 1). Within the right MTG, the difference in activation can be described as a difference in deactivation, with atypical tokens more deactivated than typical tokens. This region was compared to a similar right MTG cluster which showed sensitivity to shifts in the phonetic category boundary conditioned by lexical context producing talker-specific shifts in perception (Myers & Mesite, 2014). These clusters abutted one another, and overlapped by 3 voxels.

3.2.3. gPPI analysis

No effects of talker typicality were observed in regions typically considered core phonetic processing areas (e.g., left and right STG/STS) in the planned contrast. Nonetheless, of interest is whether the region identified as sensitive to typicality (the right MTG) is functionally connected to such regions. To this end, a generalized psychophysiological interaction (gPPI) analysis was

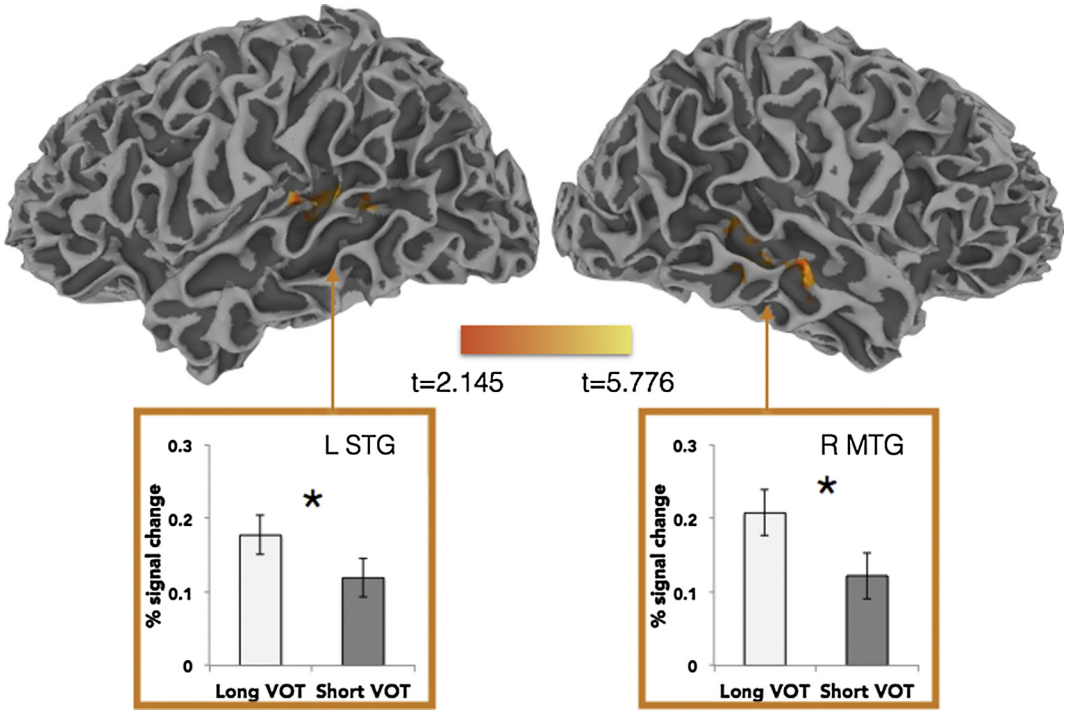

Fig. 4. Long VOT-Short VOT contrast (orange), *t*-statistic displayed. Clusters significant at $p < 0.05$, cluster corrected (voxel $p < 0.05$, minimum 196 contiguous voxels).

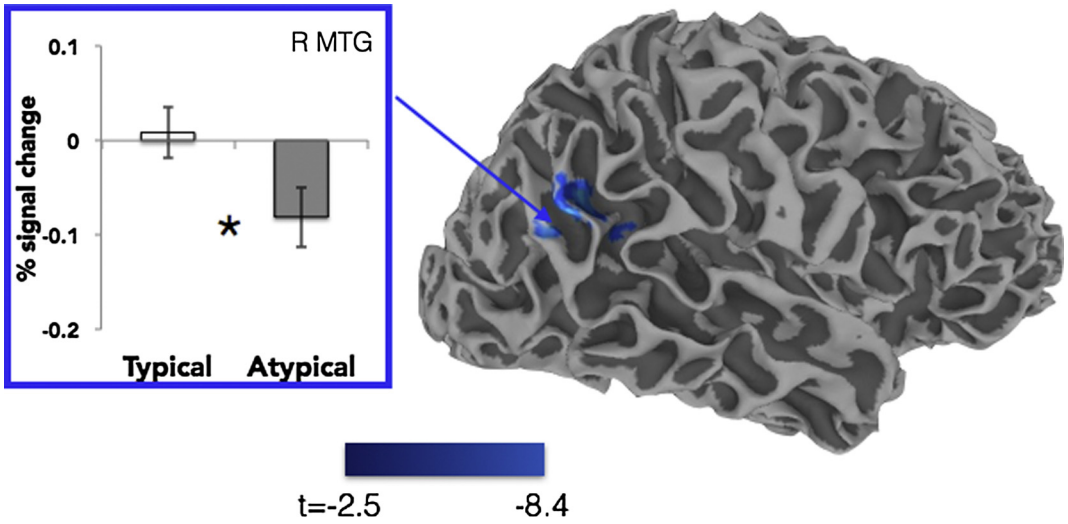

Fig. 5. Atypical-Typical contrast, *t*-statistic displayed. Clusters significant at $p < 0.05$, cluster corrected (voxel $p < 0.025$, minimum 112 contiguous voxels). Plot outlined in blue is drawn from the functional ROI located in the right middle temporal gyrus.

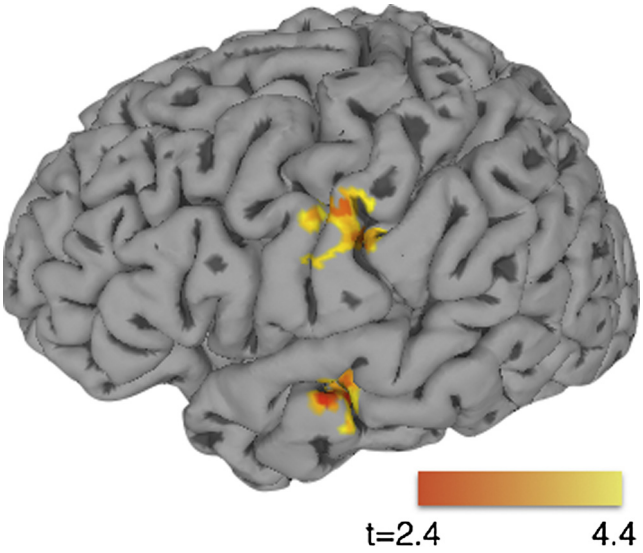

Fig. 6. Generalized psychophysiological interaction (gPPI) analysis, using the right MTG cluster identified in the Atypical > Typical contrast as a seed. Left hemisphere view shows two regions in which the magnitude of connectivity between these regions and the right MTG ROI is greater for Typical compared to Atypical tokens.

conducted using the right MTG region from the Atypical-Typical contrast as a seed (Fig. 6, Table 2). This analysis identified brain areas in which the connectivity to the seed region was modulated by the task—in this case, hearing Typical vs. Atypical trials. Two regions, one in the left postcentral gyrus extending into the precentral gyrus, and one in the left middle temporal gyrus and left superior temporal sulcus, showed significant modulation of their connectivity to the rMTG according to the trial that participants were listening to. In both cases, greater connectivity between the seed rMTG region and the area identified in the analysis was observed when participants heard Typical trials compared to Atypical trials.

4. Discussion

The speech signal carries both linguistic information and talker information. These sources of information are inextricably woven together, with often an identical acoustic cue (in this case, VOT) used to determine phonetic content as well as to inform talker identity (e.g., Theodore & Miller, 2010). Despite this, results of the current study suggest that the two uses of this same cue are separated in the neural processing stream, with left and right bilateral posterior STS responding to the 'goodness of fit' of tokens to the phonetic category (irrespective of the degree to which they are typical for a *talker*) and right temporoparietal regions responding to that same cue's typicality for a given talker (irrespective of their typicality as members of the *phonetic* category). At the same time, there is considerable cross-talk between systems for resolving talker identity and those for resolving phonetic identity—this is highlighted by the connectivity between the right MTG, which has been implicated across studies in access to talker identity (e.g., Van Lancker et al., 1989), and the left mid-to-anterior STG/STS, which has been linked to processing intelligible speech (e.g., Scott & Johnsrude, 2003). These findings add to our current understanding of the neural systems responsible for computing talker identity, suggesting that talker-specific phonetic variability, such as variations in VOT, are included as part of a talker's vocal identity representation. Of interest, despite listeners' ability to make very accurate explicit judgments about the degree to which tokens are typical or atypical for a given talker, this information does not appear to fundamentally alter the perceived 'goodness' of these tokens within areas typically associated with processing phonetic category 'goodness' (i.e., the posterior STG), which stands in contrast to some previous studies showing that talker-specific phonetic variation cascades through the processing system.

Table 2
Results from the generalized PPI analysis. Connectivity between regions below and a seed region in the RMTG for Typical > Atypical trials. Coordinates are given in Talariach space.

Cluster size in voxels	Peak x	Peak y	Peak z	Region	*t* statistic at peak
174	−55	−13	39	L. Postcentral, L. Precentral	4.248
139	−61	−10	−5	L. MTG, L. STS	4.048

4.1. Right hemisphere preference for vocal identity

A long-standing debate in the neuroscience of speech perception concerns the degree to which hemispheric laterality for speech perception is driven by differences in the acoustics of speech signal or by the functional significance of the speech signal (Boemio, Fromm, Braun, & Poeppel, 2005; McGettigan & Scott, 2012; Poeppel, 2003; Zatorre & Belin, 2001). The acoustically-driven account posits that the right hemisphere integrates acoustic information over longer-duration windows (e.g., the Asymmetric Sampling in Time hypothesis outlined in Poeppel, 2003). These longer windows are likely to map well to phonetic contrasts that are also signaled by longer, steady-state spectral information, as in the case of vowels and fricatives. In contrast, the functional view of hemispheric laterality for speech perception (see McGettigan & Scott, 2012 for review) suggests that hemispheric laterality is primarily driven by the way the speech signal is used, with phonetic processing largely left lateralized, and processing of vocal identity and other information carried by suprasegmental properties of the signal (e.g., prosody, emotion) triggering right hemisphere structures. Of interest, because previous studies of vocal identity processing have manipulated acoustic properties that are at this longer durational time frame (e.g., pitch height, formant structure), these studies have been unable to speak to the question of whether right-hemispheric recruitment for vocal identity is related to the acoustics of the signal (long duration cues) or the functional significance of those cues (cues to vocal identity). In the current study, VOT, a classic short-duration cue, was manipulated. All acoustically-based theories would place processing of this kind of short-duration cue in the left hemisphere (e.g., Poeppel, 2003; Zatorre & Belin, 2001). Nonetheless, when contrasting VOT values that are typical vs. atypical for a given talker, right hemisphere regions were still recruited. This pattern provides support for the view that right hemisphere laterality for vocal identity processing is not simply due to the types of acoustic cues that are likely to trigger right hemisphere responses, but instead that acoustic cues that are linked to vocal identity (in this case, Long or Short VOT) recruit right hemisphere structures, regardless of duration.

4.2. Talker-specific phonetic variability is part of vocal identity representations

The current results suggest that vocal identity representations in the brain are not limited to traditional *indexical* properties (e.g., F_0, jitter, prosodic contour) but also encompass variations in *phonetic* properties that are linked to the talker. Behaviorally, participants in the current study were able to explicitly judge the typicality of a VOT variant (long or short) as typical or atypical of a given talker with near-ceiling accuracy. Even when tasked with a phonetic categorization decision that did not require attention to talker typicality, regions in the right MTG/supramarginal gyrus (SMG) showed sensitivity to the typicality of these variants. Indeed, it is striking that VOT typicality was encoded at all, given that there was ample acoustic variation in the signal (e.g., F_0 values, prosody) to separate the two talker's voices. Recall that the typicality analysis collapsed both Long and Short VOT variants for both talkers (each of which might be perceived as 'typical' or 'atypical' of a given talker according to the counterbalanced group assignment), so differences in activation seen within this contrast cannot be attributed to surface-level properties of the stimuli. Notably, right posterior temporal and parietal regions have been linked in lesion studies (Van Lancker et al., 1988, 1989) and imaging studies (Andics et al., 2010, 2013) to access to talker identity. This region also abuts a slightly more ventral MTG area that was found to be responsive in a previous study to talker-specific phonetic variability when that variability takes the form of an ambiguous phoneme inserted in a biasing lexical context (Myers & Mesite, 2014). Taken together, this suggests that the right posterior temporoparietal area that responds to shifts in the phonetic category boundary when those shifts are triggered by talker-specific phonetic variability (Myers & Mesite, 2014) also responds to talker-specific phonetic variability when *no* shift in the phonetic category boundary is required. Previous studies of talker identity processing have often focused on examining neural responses in reaction to changing vocal identity while holding the linguistic context constant (Salvata et al., 2012; Wong et al., 2004), morphing the signal to systematically manipulate all of the acoustic cues associated with a talker (e.g., Andics et al., 2010), or manipulating acoustic properties such as formant structure that trigger a change talker identity (e.g., Kreitewolf, Gaudrain, & von Kriegstein, 2014; von Kriegstein et al., 2010). However, it is unclear whether sensitivity to talker information in these studies is related to (potentially automatic) detection of talker-specific properties that necessarily vary along with changing vocal tract anatomy. For instance, changes in the overall formant structure of a talker are fairly deterministically related to the length of the vocal tract, and changes in the mean F_0 of a speaker are closely related to the mass of the vocal folds. The current results expand on this literature to explicitly explore the role that specific acoustic-phonetic properties play in neural processing for talker identity when those cues are part of the talker idiolect. Specifically, the current results suggest that within-category phonetic variation (together with vocal pitch, timbre, prosody, and other suprasegmental cues) form part of the talker idiolect representation and are quickly integrated into the neural processing stream for vocal identity. This view converges with evidence suggesting that phonetic exposure and expertise influences voice recognition, for instance, facilitating identification of voices in one's own language compared to non-native languages (Kadam, Orena, Theodore, & Polka, 2016; Orena et al., 2015).

Of interest, the neural talker typicality effect was accompanied by behavioral evidence suggesting that talker-atypical tokens were slightly more difficult to process than talker-typical tokens. This was evident in the reaction time analysis, which showed slowed responses to tokens that were atypical of the talker, even as participants performed a phonetic categorization task that did not require them to consider the talker at all. This is perhaps surprising given that the VOT values chosen for this study fell squarely within the range of standard voiceless tokens—that is, there was little to no actual ambiguity in these stimuli with respect to their phonetic category identity. This behavioral interaction underlines the co-dependence of talker information and phonetic information, and suggests that talker effects are in some sense unavoidable as listeners map acoustic-phonetic details to phonetic categories. It also raises a potential interpretation of our results in that the right MTG cluster may respond to difficulty of the stimuli in general, rather than talker typicality per se. However, it is worthy to note that no such activation was observed in other areas often implicated in executive processing (e.g., IFG, cingulate, see Badre, 2013 for review), and as such, we interpret the right MTG finding as convergent with a large literature implicating these regions in talker processing specifically.

4.3. Talker influences on phonetic processing

Thus far, evidence suggests that phonetic information pervades the representation of vocal identity in the brain. What is less clear is whether vocal identity similarly pervades phonetic processing in the brain. As reviewed above, the bilateral posterior STG respond to the 'goodness of fit' of tokens to their phonetic category (Myers, 2007). Relevant for the current results, along a VOT continuum, VOTs that are 'typical' for English talkers require less activation than those that are atypical, a pattern that is thought to reflect

the ease of processing the more prototypical phonetic tokens. That pattern was replicated in the current study, in that Long VOT tokens (which are atypical in normal input) showed greater activation than Short VOT tokens (which are closer to typical values for natural speech). While it is possible that low-level differences (e.g., length, spectral energy distributions) between stimuli might drive this activation difference, the finding here nonetheless closely replicates previous results linking such a pattern to goodness of fit specifically (Myers, 2007). Of interest, previous work demonstrates that phonetic sensitivity in the temporal lobes *can* be modulated by context; specifically, top-down influences from lexical status and sentential bias have been shown to shift phonetic sensitivity in the temporal lobes (Gow et al., 2008; Guediche et al., 2013; Myers & Blumstein, 2008). However, this posterior STG system showed no such modulation in the current study.

This apparent contradiction points to a potential qualitative distinction in how linguistic and talker information modulate phonetic processing. An important contrast is that talker identity is not required for resolving phonetic ambiguity in the current stimulus set; all tokens were unambiguous members of their phonetic category. In this sense, while listeners may track typical and atypical variants, no change in criteria is needed to facilitate phonetic processing. In contrast, the above-cited studies have all demonstrated top-down shifts in phonetic categorization (and concomitant shifts in temporal lobe sensitivity) for *ambiguous* input. That is, top-down information is actually required in order to resolve category membership and facilitate processing. We speculate that when talker information is such that resolution of talker identity also impacts resolution of phonetic category identity (for instance, in the case of a non-native accent), shifts in temporal lobe sensitivity will be evident.

While the regions that responded to phonetic category typicality as representative of the language (Long vs. Short VOT) were not affected by talker typicality when measured in terms of magnitude of activation, modulation of left temporal regions was seen in the connectivity analysis. Specifically, here a right-hemisphere MTG region that was sensitive to talker typicality showed connectivity to an anterior STG cluster that increased when participants were listening to talker-typical as compared to talker-atypical stimuli. This finding closely resembles results from Kriegstein and Giraud (2004) and von Kriegstein et al. (2010), in which increased right posterior temporal-to-left anterior temporal connectivity was observed as the talker varied. Collectively, these results are consistent with the interpretation that right hemisphere regions are responsible for aggregating information about the talker, including information that can be used to infer anatomical variations (formant structure, F0) but also information that is probabilistically related to the talker's idiolect (talker differences in VOT). Importantly, all of these cues are useful for phonetic processing—formant structure determines vocal tract length as well as vowel identity, fundamental frequency cues vocal fold mass as well as lexical tone contour, and VOT can be related to a talker, but is also diagnostic of word-initial stop voicing. The profile of talker acoustics in the right MTG subsequently informs linguistic processing of stimuli in the left temporal lobes. This general view, which closely resembles one advanced in Kreitewolf et al. (2014), posits that left-to-right connectivity waxes and wanes according to the demands of the task (e.g., focusing on linguistic properties compared to talker properties), and the familiarity of the talker (e.g., varying according to the strength or reliability of the talker-specific voice profile built in the right hemisphere).

5. Conclusion

Behavioral evidence suggests that talker-specific phonetic variation is mutually used for talker identification and phonetic processing; namely, associating phonetic variation with a given talker facilitates processing of the content of speech (e.g., Nygaard et al., 1994) and facilitates talker identification (e.g., Goggin et al., 1991). Current brain data suggest that despite this tight link in behavior, phonetic processing and voice recognition streams are separately sensitive to within-category phonetic variation, but that these streams re-integrate in cross-hemispheric connections that are modulated by talker typicality. Future work should consider how relationships between these networks change over the time-course of learning, as listeners accumulate substantial experience with the phonetic detail of a talker's voice. That is, it remains to be seen whether sensitivity to talker-specific phonetic detail will ultimately emerge in left superior temporal regions that have previously shown adaptation as a consequence of top-down (lexical) feedback. If not, then the previously-observed benefits for talker familiarity on comprehension must arise from a different source, perhaps the recruitment of right-lateralized talker systems during comprehension. Future research in this vein will contribute towards a neurobiological account of talker-familiarity effects that are ubiquitous in the language comprehension literature.

Authors' contribution

Both EBM and RMT contributed to the study design and were responsible for data collection. EBM analyzed the MRI data and RMT analyzed the behavioral data. Both authors were involved in preparing the manuscript for submission.

Acknowledgments

Gratitude is extended to Laura Mesite for her assistance with experiment programming and to Sahil Luthra for his assistance with data collection. Preliminary data were presented at the 2015 annual meeting of Cognitive Neuroscience Society (Chicago, Illinois). This research was supported by NIH NIDCD grants R03 DC009395 and R01 DC013064 to EBM. This content is the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDCD.

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.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113*(1), 544–552.

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

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage, 52*(4), 1528–1540. http://dx.doi.org/10.1016/j.neuroimage.2010.05.048.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*(3), 163–187.

Badre, D. (2013). Hierarchical cognitive control and the functional organization of the Frontal Cortex. *The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges, 2*, 300.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105–2109.

Blumstein, S. E., & Myers, E. B. (2014). Neural systems underlying speech perception. In K. Oshsner & S. Kosslyn (Eds.). *Oxford handbook of cognitive neuroscience* (Vol. 2). .

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*(3), 389–395. http://dx.doi.org/10.1038/nn1409.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences, 11*(12), 535–543. http://dx.doi.org/10.1016/j.tics.2007.10.001.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience, 13*(11), 1428–1432. http://dx.doi.org/10.1038/nn.2641.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647. http://dx.doi.org/10.1121/1.1815131.

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

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass, 5*(5), 190–204. http://dx.doi.org/10.1111/j.1749-818X.2011.00276.x.

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.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224.

Gelfer, M. P. (1988). Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice, 2*(4), 320–326. http://dx.doi.org/10.1016/S0892-1997(88)80024-9.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448–458. http://dx.doi.org/10.3758/BF03199567.

Gow, D. W., Segawa, J. A., Ahlfors, S. P., & Lin, F.-H. (2008). Lexical influences on speech perception: A Granger causality analysis of MEG and EEG source estimates. *NeuroImage, 43*(3), 614–623. http://dx.doi.org/10.1016/j.neuroimage.2008.07.027.

Guediche, S., Salvata, C., & Blumstein, S. E. (2013). Temporal cortex reflects effects of sentence context on phonetic processing. *Journal of Cognitive Neuroscience, 25*(5), 706–718. http://dx.doi.org/10.1162/jocn_a_00351.

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.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science, 14*(5), 1002–1011. http://dx.doi.org/10.1111/j.1467-7687.2011.01052.x.

Kadam, M. A., Orena, A. J., Theodore, R. M., & Polka, L. (2016). Reading ability influences native and non-native voice recognition, even for unimpaired readers. *Journal of the Acoustical Society of America, 139*(1), EL6–EL12. http://dx.doi.org/10.1121/1.4937488.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15.

Kreitewolf, J., Gaudrain, E., & von Kriegstein, K. (2014). A neural mechanism for recognizing speech spoken by different speakers. *NeuroImage, 91*, 375–385. http://dx.doi.org/10.1016/j.neuroimage.2014.01.005.

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

Liebenthal, E., Binder, J. R., Spitzer, S. M., Possing, E. T., & Medler, D. A. (2005). Neural substrates of phonemic perception. *Cerebral Cortex, 15*(10), 1621–1631.

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. http://dx.doi.org/10.1016/j.tics.2012.04.006.

McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): A comparison to standard approaches. *NeuroImage, 61*(4), 1277–1286. http://dx.doi.org/10.1016/j.neuroimage.2012.03.068.

McMurray, B., Tanenhaus, M. K., Aslin, R. N., & Spivey, M. J. (2003). Probabilistic constraint satisfaction at the lexical/phonetic interface: Evidence for gradient effects of within-category VOT on lexical access. *Journal of Psycholinguistic Research, 32*(1), 77–97.

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.

Myers, E. B., & Blumstein, S. E. (2008). The neural bases of the lexical effect: An fMRI investigation. *Cerebral Cortex, 18*(2), 278–288.

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. doi: PSCI2380 [pii] 10.1111/j.1467-9280.2009.02380.x.

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. http://dx.doi.org/10.1016/j.jml.2014.06.007.

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.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376.

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

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition, 143*, 36–40. http://dx.doi.org/10.1016/j.cognition.2015.06.002.

Perrachione, T. K., & Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*(8), 1899–1910.

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.

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics, 15*(2), 285–290.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time". *Speech Communication, 41*(1), 245–255. http://dx.doi.org/10.1016/S0167-6393(02)00107-3.

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

Salvata, C., Blumstein, S. E., & Myers, E. B. (2012). Speaker invariance for phonetic information: An fMRI investigation. *Language and Cognitive Processes, 27*(2), 210–230. http://dx.doi.org/10.1080/01690965.2011.594372.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences, 26*(2), 100–107.

Stevens, A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research, 18*, 162–171.

Talairach, J., & Tournoux, P. (1988). *A co-planar stereotaxic atlas of a human brain.* Stuttgart: Thieme.

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. http://dx.doi.org/10.1121/1.3467771.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America, 125*(6), 3974–3982.

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *Journal of the Acoustical Society of America, 138*(2), 1068. http://dx.doi.org/10.1121/1.4927489.

Utman, J. A., Blumstein, S. E., & Sullivan, K. (2001). Mapping from sound to meaning: Reduced lexical activation in Broca's aphasics. *Brain and Language, 79*(3), 444–472. http://dx.doi.org/10.1006/brln.2001.2500.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex, 24*(2), 195–209.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*(5), 665–674.

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. *Brain Research. Cognitive Brain Research, 17*(1), 48–55.

von Kriegstein, K., Smith, D. R. R., Patterson, R. D., Ives, D. T., & Griffiths, T. D. (2007). Neural representation of auditory size in the human voice and in sounds from other resonant sources. *Current Biology: CB, 17*(13), 1123–1128. http://dx.doi.org/10.1016/j.cub.2007.05.061.

von Kriegstein, K., Smith, D. 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.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*(7), 1173–1184. http://dx.doi.org/10.1162/0898929041920522.

Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *Journal of the Acoustical Society of America, 137*(1), 419. http://dx.doi.org/10.1121/1.4904699.

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11*(10), 946–953.