Published in final edited form as:
Lang Cogn Process. 2012 ; 27(2): 210–230. doi:10.1080/01690965.2011.594372.

Speaker Invariance for Phonetic Information: an fMRI Investigation

Caden Salvata[1], **Sheila E. Blumstein**[1], and **Emily B. Myers**[1,2]
[1]Brown University
[2]University of Connecticut

Abstract

The current study explored how listeners map the variable acoustic input onto a common sound structure representation while being able to retain phonetic detail to distinguish among the identity of talkers. An adaptation paradigm was utilized to examine areas which showed an equal neural response (equal release from adaptation) to phonetic change when spoken by the same speaker and when spoken by two different speakers, and insensitivity (failure to show release from adaptation) when the same phonetic input was spoken by a different speaker. Neural areas which showed speaker invariance were located in the anterior portion of the middle superior temporal gyrus bilaterally. These findings provide support for the view that speaker normalization processes allow for the translation of a variable speech input to a common abstract sound structure. That this process appears to occur early in the processing stream, recruiting temporal structures, suggests that this mapping takes place prelexically, before sound structure input is mapped on to lexical representations.

Introduction

One of the unsolved problems in speech perception is the “invariance problem” – how a listener perceives a stable phonetic category despite variation in the acoustic signal. One critical source of variability in the speech stream comes from speech produced by different talkers. Owing to different vocal tract sizes and to different dialectal and idiosyncratic characteristics, the acoustic output for the same phonetic category, e.g. [d], or the same word, e.g. ‘dog’ produced by different talkers is not the same. And yet, listeners are able to map this variable acoustic input on to the same lexical form (sound structure) and ultimately conceptual representation. At the same time, listeners are also able to distinguish and identify different speakers, despite the fact that the message may be phonetically varied. Thus, listeners must be able to map the variable acoustic input on to a common sound structure representation while being able to retain phonetic detail to distinguish among the identity of talkers.

There are a number of theories which have attempted to solve the invariance problem. Abstractionist theories of speech recognition assume that a perceptual normalization process matches acoustically variable stimuli onto fixed abstract mental representations (Stevens, 1960). In this process, variability in the acoustic signal including speaker information is considered to be ‘noise’ and is filtered out from the processing stream so that the idealized stimuli may be mapped on to an abstract phonetic representation (Nearey, 1989). However, a series of studies using a broad range of behavioral measures has shown that information

Correspondence address: Sheila E. Blumstein, Brown University, Department of Cognitive, Linguistic, and Psychological Sciences, 229 Waterman Street, Providence, RI 02912, Sheila_Blumstein@brown.edu.

about a speaker's voice affects perception and memory for speech, suggesting that such indexical features are perceived, encoded, and used by the listener in the processes of word recognition and recall (Goldinger, 1997). This evidence has led to the episodic theory of speech perception, in which fine details of the speech input including speaker information are preserved in memory and together form the substrate for phonetic categories and ultimately lexical representations. In this view, the variability inherent in the input provides critical information that is both used and retained by the listener in processing speech (Johnson & Mullennix, 1997).

The abstractionist and episodic models of speech perception differ concerning their predictions about speaker invariance, defined here as selective sensitivity to the phonetic properties of speech and insensitivity to speaker identity. In the abstractionist model, the system must achieve speaker invariance as a necessary step in the perceptual normalization of the speech signal, since such indexical information is 'irrelevant' to the underlying phonetic code and hence must be discarded. In the episodic model, however, speaker invariance is not required for speech perception. In fact, phonetic details, including speaker-specific information, form a part of the representation of both sounds and ultimately words.

The goal of the present study is to investigate whether there are neural areas that show speaker invariance in the context of sensitivity to phonetic processing. In particular, we examined whether there are neural areas that show similar neural responses to phonetic change irrespective of whether the phonetic change is spoken by the same or different speakers.

Lesion studies have shown a double dissociation between speech and speaker processing abilities. Phonagnosia, a disorder identified by Van Lancker and Canter (1982), is characterized by difficulty in recognizing familiar voices and discriminating between unfamiliar ones, and is associated with right hemisphere temporal and parietal damage. Case studies with phonagnosic patients have found that speech perception can be retained despite the loss of speaker recognition and impaired speaker discrimination (Peretz, Kolinsky, Tramo, Labreque, Hublet et al., 1994). In contrast, aphasic patients with left hemisphere lesions appear to have intact speaker recognition in the presence of speech perception impairments (Van Lancker, Kreiman, & Cummings, 1989). This constellation of deficits suggests that phonetic and speaker processing are at least partly dissociable in the cortex. However, they do not show whether speaker information is 'filtered out' from phonetic information during speech perception and word recognition processes.

Consistent with the lesion evidence, it has been shown that the right anterior superior temporal sulcus (STS) is selectively recruited in speaker identity (Belin & Zatorre, 2003) and the superior temporal sulcus (STS) plays a role in processing human voice information (Belin & Zatorre, 2003; von Kriegstein & Giraud, 2004; see Belin, Fecteau, & Bedard, 2004 for review). Of interest, homologous areas of the temporal lobe appear to be activated for both speaker and speech information with greater right hemisphere lateralization for speaker and greater left hemisphere activation for speech information. In particular, von Kriegstein et al. (2003) showed that given the same auditory input, directing attention to voices activated the right middle STS and directing attention to verbal content activated the left middle STS.

Nonetheless, in contrast to the lesion literature, other fMRI studies suggest that the right hemisphere specialization for speaker identity and left hemisphere specialization for phonetic content may be relative rather than absolute. Both speech as well as non-speech vocalizations have been shown to significantly activate regions in the mid and posterior bilateral STS, including vocalizations from other species (Belin, Zatorre, Lafaille, Ahad, &

Pike, 2000; Fecteau, Armony, Joanette, & Belin, 2004; Warren, Scott, Price, & Griffiths, 2006). Additionally, a bilateral temporal-parietal network including the middle temporal gyrus (MTG) as well as superior parietal areas has been shown to be recruited when listeners are required to attend to target words spoken by different speakers, (Wong, Nusbaum, & Small, 2004). Neuroimaging studies looking specifically at the perception of the phonetic categories of speech within the same speaker or vocal tract have identified similar regions: the superior temporal cortex has been implicated in early phonetic analysis (Blumstein, Myers, & Rissman, 2005; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Binder & Price, 2001), while temporal and parietal structures such as the middle temporal gyrus (MTG), angular gyrus (AG) and supramarginal gyrus (SMG) have shown sensitivity to phonetic category differences (Celsis, Boulanouar, Doyon, & Ranjeva,1999; Joanisse, Zevin, & McCandliss, 2007; Zevin & McCandliss, 2005).

Several recent studies examining both perception of phonetic properties of speech as well as speaker identity suggest that speech and voice processing may share neural resources and may be processed by similar mechanisms. Using a same/different speech or loudness control task, von Kriegstein and colleagues showed that left posterior STG/STS encodes both speaker-related vocal tract parameters and speech (von Kriegstein, Smith, Patterson, Kiebel, & Griffiths, 2010). In particular, vocal tract length parameters which influence both speaker recognition processes and phonetic properties of speech activate the left STG/STS, and the degree of activation was modulated by a speech discrimination task. Using multivariate pattern recognition, Formisano, De Martino, Bonte, & Goebel (2008) showed that the STS bilaterally is responsive in a distributed fashion to vowel identity independent of speaker and speaker identity independent of vowel (Formisano et al., 2008). And Leaver and Rauschecker (2010) showed that while bilateral STS/STG clusters were shown to be selective to human speech generally, a subregion within the left STS/STG cluster was sensitive to acoustic-phonetic content. Taken together, these studies suggest that the left hemisphere may play a role in integrating the two sources of information critical for language communication, speaker and phonetic information. What is not clear from these studies is whether there is a neural region that shows speaker invariance for phonetic information.

We report an fMRI experiment in which the adaptation paradigm was utilized to examine whether neural areas may be identified that show speaker invariance. This paradigm is based on physiological studies of adaptation in which the repeated presentation of a stimulus fatigues a neuron or population of neurons. The consequence of such repeated presentations using fMRI is a reduction in the BOLD response. The subsequent presentation of a stimulus differing across a relevant dimension results in a signal increase or a release from adaptation in areas sensitive to that dimension (Grill-Spector & Malach, 2001), although cf. Zevin, Yang, Skipper, & McCandliss, 2010 for an alternative explanation for the increased activation observed for ‘different’ trials). This technique has been previously used to investigate the neural systems underlying phonetic category perception. For example, Celsis et al. (1999) and Zevin and McCandliss (2005) both showed dishabituation (or release from adaptation) to different naturally produced syllables in the posterior STG near the border of the SMG. Using a place of articulation continuum, Joanisse et al. (2007) showed greater release from adaptation for between than within category changes in left STS and MTG and inferior parietal cortex involving the AG and SMG.

Myers, Blumstein, Walsh, & Eliassen (2009) used the adaptation paradigm to examine those brain regions in the phonetic processing stream that showed phonetic category invariance. In particular, they attempted to identify neural areas that responded differently to between phonetic category changes but similarly to within phonetic category changes in the perception of voice-onset time (VOT), an acoustic parameter that distinguishes voiced and

voiceless stop consonants. Subjects were presented with trains of five phonetic tokens, in which the first four tokens were identical. In two experimental conditions the fifth token was varied – either it belonged to the same phonetic category but possessed a different VOT (within phonetic category change), or it belonged to a different phonetic category (between phonetic category change). Several areas (bilateral IFG, left posterior STG, and left MTG) were found to be sensitive to between phonetic category change. However, only one of these regions, a cluster in the left inferior frontal sulcus, showed phonetic category invariance, i.e. insensitivity to within category change. This pattern of selective sensitivity suggests that the IFG is involved in the categorical perception of speech, and plays a role in computing categorical representations for speech.

Previous findings in visual category perception in both monkeys and humans have also shown that the IFG is involved in achieving category invariance for visual objects, supporting the hypothesis that this region plays a domain-general role in the computation of category representations (Freedman, Riesenhuber, Poggio, & Miller, 2002, 2003; Jiang, Bradley, Rini, Zeffiro, VanMeter et al., 2007). What is not clear is whether the same neural resources and computational principles that are used for deriving phonetic category invariance are also used for deriving speaker invariance, i.e. phonetic category constancy across speaker variability. If the two sources of variability (speaker and phonetic) are treated equivalently, then an invariant neural response should emerge in the left IFG. However, as described above, prior neuroimaging work has suggested that speaker information is extracted earlier in the processing stream, recruiting the STS/STG bilaterally. Given that speech processing recruits primarily the left hemisphere, a likely candidate area for speaker invariance would be the left STS/STG.

Materials and Methods

Subjects

Eighteen right-handed healthy participants who reported normal hearing and no history of neurological disorders (15 women, mean age 22.61±4.82) were recruited from the staff and students of Brown University. Handedness was confirmed using the Edinburgh Inventory (Oldfield, 1971) with a mean laterality quotient of 82.04±17.85. All subjects were screened for MR safety prior to scanning. Subjects gave written informed consent in accordance with the Human Subjects Policies of Brown University and the Helsinki Declaration of 1975, as revised in 1983, and were paid for their participation. All subjects participated in a behavioral pretest several days prior to being scanned.

Stimuli

Stimuli consisted of the phonetic tokens [ga] and [ta] each produced by a male and female speaker of American English. To obtain these materials, speakers were instructed to repeat each token several times. Four tokens were selected, two [ga] tokens and two [ta] tokens produced by each speaker. These tokens had similar pitch contours as measured by the BLISS software system (Mertus, 2000), were edited to 350msec duration using Praat (Boersma, 2001), and were adjusted to equivalent volume by ear. These stimuli were used for both the behavioral pretest and the imaging experiment. Four additional stimuli were created to be used infor the fMRI experiment. These stimuli consisted of the [ga] and [ta] tokens with the volume of each stimulus lowered by 85%.

Behavioral Pretest

A behavioral pretest was conducted to assess whether the stimulus conditions to be used in the fMRI experiment differed in processing difficulty. Stimuli were presented in pairs belonging to one of the following conditions: *Same*, i.e. $[ta]_{S1}$ $[ta]_{S1}$, in which the same

token was presented twice, *Phonetic Change*, i.e. $[ta]_{S1}$ $[ga]_{S1}$, in which the phonetic category changed but speaker remained the same, *Speaker Change*, i.e. $[ta]_{S1}$ $[ta]_{S2}$, in which phonetic category remained the same but speaker identity changed, and *Both Change*, i.e. $[ta]_{S1}$ $[ga]_{S2}$. Subjects performed a same/different judgment task in which they were instructed to attend to either speaker identity or phonetic category and determine by button press whether the two syllables shared or did not share the designated characteristic. The pretest consisted of four blocks, each consisting of 32 trials equally divided among the four conditions and pseudo-randomized. In alternate blocks subjects were asked to judge whether the two syllables were spoken by the same speaker (Speaker Task) or whether the two syllables were phonetically the same (Phonetic Task). Presentation of the blocks was counter-balanced across subjects. Accuracy and reaction times were recorded.

Results were averaged across subjects by condition (Speaker Change, Phonetic Change, Both Change) and response type (same or different). Table 1 shows the results of performance accuracy and reaction-time latencies for correct different responses. As the Table shows, performance accuracy was very high ranging between 99 and 100% across conditions. A one-way ANOVA for reaction-time was significant ($F(3, 51) = 4.544$, $p<.007$). Post-hoc tests revealed that subjects were significantly faster at responding to Speaker Change than they were to responding to Phonetic Change ($p = .025$). No significant differences were found between the phonetic task and the speaker task when both phonetic and speaker dimensions changed ($p=.072$).

fMRI Experiment

Each trial consisted of five tokens separated by a 50 msec ISI. There were four experimental conditions, paralleling the pretest. In the *Adaptation* condition, all five stimuli were identical. In the *Phonetic Change* condition, four identical stimuli were followed by a stimulus differing only in phonetic category, i.e. $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ga]_{S1}$]. In the *Speaker Change* condition, four identical stimuli were followed by a stimulus differing only in speaker identity, i.e. $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S2}$. In the *Both Change* condition, four identical stimuli were followed by a stimulus differing in both speaker identity and phonetic category, i.e. $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ta]_{S1}$ $[ga]_{S2}$. In addition, in *Target* trials, a low-volume target stimulus appeared within the trial. These stimuli occurred equally in trials of each experimental type. Subjects were instructed to respond with a button press using their right hand when they heard a stimulus of lower volume.

There were a total of 36 trials per condition as well as 36 low-volume target trials, 9 in each of the experimental conditions. Stimuli were presented to the subjects binaurally through commercially available pneumatic headphones using Bliss fMRI Runner (www.mertus.org).

Data Acquisition

MRI data were collected using a 3T Siemens Trio scanner with a standard 8-channel head coil. For each subject, high-resolution T1 weighted structural images were acquired for anatomical co-registration (TR=1900ms, TE=4.15ms, TI=1100ms, $1mm^3$ isotropic voxels, 256 × 256 matrix). Functional images were acquired with an echo-planar sequence (TR=1sec, TE=30ms, flip angle = 90 degrees, FIV=192mm3, 64 × 64 matrix) in 15 5mm-thick ascending interleaved axial slices. EPI acquisition used a slow event-related, clustered acquisition design, with stimuli presented during a 2.3sec silent gap between functional scans and a 14.2sec intertrial interval to prevent hemodynamic overlap. Each run consisted of 228 volumes, for a total of 912 in the experiment. Each subject completed all four runs, with the exception of one subject who only completed three.

Data Analysis

Functional data were analyzed using AFNI (Cox, 1996). The first four volumes from each run were censored to prevent T1 saturation effects. Functional data sets were motion-corrected using a six-parameter rigid body transform. Two runs were discarded, one due to excessive movement in one participant, and another due to stimulus presentation error in another participant. The MPRAGE anatomical scan for each subject was normalized to Talairach and Tournoux stereotaxic space. The functional data were then aligned to the anatomical scan, resampled to $3mm^3$, spatially smoothed with a 6 mm Gaussian kernel and converted to percent signal change units.

Timing files were created for each condition and convolved with a gamma function to model the idealized hemodynamic response. Preprocessed BOLD data were then masked using a sixteen-subject composite mask and submitted to a regression analysis with the idealized waveforms as regressors. The six parameters from the motion-correction process were included as nuisance regressors, as were baseline, linear and quadratic trends.

A mixed-factor ANOVA was performed with subjects as a random factor and stimulus condition as a fixed factor. A mask consisting of areas previously implicated in language processing (AG, IFG, MFG, MTG, SMG, STG) was used as the basis of a small volume correction in which the following three planned comparisons were performed: Phonetic Change vs. Adaptation which allowed for identification of clusters sensitive to phonetic change when spoken by the same speaker, Speaker Change vs. Adaptation which allowed for identification of clusters sensitive to speaker change when the same syllable was produced, and (Both Change + Phonetic Change) vs. (Speaker Change + Adaptation) which allowed for identification of clusters showing sensitivity to phonetic change regardless of speaker change-related activation. All comparisons were thresholded to $p<0.05$, corrected for multiple comparisons (minimum cluster determined by Monte Carlo simulation, 33 contiguous voxels at a voxel-wise $p<0.025$).

Results

Regions sensitive to phonetic change

Activations for the Phonetic Change > Adaptation contrast are shown in Table 2, with group analysis t-statistics displayed in Figure 1. Clusters emerged in the STG bilaterally with the largest STG cluster on the left and in BA44 in the left IFG.

Regions sensitive to speaker change

Activation for the Speaker Change > Adaptation contrast is shown in Table 2, with group analysis activation displayed in Figure 1. A temporal lobe cluster emerged in this contrast, located in the left posterior STG.

Speaker invariant regions

The Phonetic plus Both Change > Speaker Change plus Adaptation contrast revealed four clusters in the temporal lobe including the STG bilaterally, the left MTG and left temporal pole and one in the left prefrontal cortex (see Table 2). The STG clusters were particularly large (235 voxels on the left, 191 on the right). Based on fMRI results suggesting that the anterior and posterior STG are functionally distinct (Binder, Frost, Hammeke, Bellgowan, Springer et al., 2000; Britton, Blumstein, Myers, & Grindrod, 2009; Davis & Johnsrude, 2003; Giraud, Kell, Thierfelder, Sterzer, Russ et al., 2004; Scott, Blank, Rosen, & Wise, 2000; Hickock & Poeppel, 2004; Scott & Wise, 2004), both left and right STG clusters were divided in half along the y-axis. This resulted in a total of seven clusters for this contrast. The pattern of results for these clusters indicates areas sensitive to phonetic change

irrespective of speaker. However, they do not show whether these areas show speaker invariance, that is, an identical neural response to phonetic change whether speaker changes or not.

In order to identify regions showing speaker invariance, these seven clusters were submitted to a region of interest analysis in which mean percent signal change for each condition was calculated for each subject. A 2 × 2 ANOVA was run on these data and comparisons across conditions were submitted to post-hoc Student-Newman Keuls tests. Speaker invariance was defined as those clusters showing a release from adaptation for both Phonetic Change and Both Change conditions with no difference in the activation patterns between them (phonetic category constancy across speaker variability) and no release from adaptation for the Speaker Change condition (insensitivity to speaker change alone). Only two regions, which were located in the anterior STG bilaterally, showed this pattern (see Figure 3).

Discussion

The goal of the current study was to examine whether any neural areas could be identified that showed speaker invariance. To this end, an adaptation paradigm was utilized to examine areas that showed an equal neural response (equal release from adaptation) to phonetic change when spoken by the same speaker and when spoken by two different speakers. These areas also needed to show insensitivity (failure to show release from adaptation) for the same phonetic input when spoken by a different speaker. Results showed speaker invariance in the anterior superior temporal gyrus bilaterally. In addition, results replicated earlier work showing sensitivity to changes in phonetic category in the STG and IFG and changes in speaker identity in the left posterior superior temporal gyrus.

Phonetic Category Change Sensitivity

The current study showed sensitivity to phonetic category change in both bilateral STG and left IFG. These results replicate previous findings showing temporal lobe and prefrontal sensitivity to phonetic category change (Scott & Wise 2004; Blumstein et al., 2005; Myers et al., 2009). The left STG cluster that emerged in the current study partially overlapped (40 out of 249 voxels) that of Myers et al. (2009). The left IFG cluster revealed in the contrast falls within the frontal region found by Myers et al. (2009) showing phonetic category sensitivity. In contrast to the Myers et al. study, however, the current study failed to find any significant clusters in the right IFG.

Zevin and colleagues (Experiment 2, 2010), reported results from a study which, like the present study, used an adaptation design to investigate neural responses to speaker change and phonetic change. Consistent with the results of the present study, they reported activation in superior temporal regions for both speaker and phonetic change. Unlike the present study, however, they did not show frontal activation for phonetic category change. This discrepancy may be due to differences in imaging coverage; Zevin and colleagues focused on temporo-parietal areas and only partially imaged frontal areas. Moreover, in the condition in which there was no phonetic change, Zevin et al. used different utterances belonging to the same phonetic category spoken by the same speaker. It is possible that the use of variable ‘same’ stimuli may have resulted in a smaller reduction in activation compared to repetition of identical stimuli, given that both the IFG and STG have shown sensitivity to within phonetic category differences (Myers et al., 2009).

Speaker Change Sensitivity

Sensitivity to speaker change was observed in the left posterior STG. These findings are consistent with several studies showing left STS/STG activation when focusing on the

encoding of speaker-related vocal tract parameters (von Kriegstein, Warren, Ives, Patterson & Griffiths, 2006; von Kriegstein, Smith, Patterson, Ives, & Griffiths, 2007; von Kriegstein et al., 2010) and when accessing words spoken by multiple speakers (Wong et al., 2004). Nonetheless, most experiments investigating speaker change have shown bilateral posterior STS/STG activation with a tendency for right hemisphere lateralization (see Belin et al., 2004 for review). And Belin and Zatorre (2003), using a voice selective adaptation paradigm suggested that the right anterior STS/temporal pole is a speaker-sensitive region. No activation was shown in this latter area in the current study because of signal dropout. However, this does not explain why the current study failed to show any sensitivity to speaker change in the right hemisphere. Results from the behavioral pretest suggest one possibility. It will be recalled that subjects were significantly faster responding to speaker change than they were to responding to phonetic change suggesting that the speaker change task was easier than the phonetic change task (Mullenix & Pisoni, 1990). Although the fMRI experiment did not require participants to direct attention to either speaker change or phonetic change, the results of the behavioral pretest suggest that changes in speaker required fewer neural resources than changes in phonetic category resulting in overall less activation in this condition. Our neuroimaging results are consistent with the behavioral results. When the cluster threshold for the Speaker vs Adaptation contrast was reduced to $p<.05$, a 21 voxel cluster emerged in the right posterior middle temporal gyrus (38, −74, 24) and a 13 voxel cluster emerged in the right mid-STG (55, −13, 6), indicating that the right hemisphere was recruited although to a substantially weaker threshold.

Speaker Invariance

Seven clusters exhibited greater activation for the conditions in which there was a phonetic category change (Phonetic Change and Both Change) compared to the two conditions in which phonetic category was constant (Adaptation and Speaker Change). Of these seven, only bilateral activation in the anterior STG exhibited the invariance pattern, showing significant activation in response to Phonetic Change and Both Change, an equal neural response to Phonetic Change and Both Change, and no difference between the Speaker Change and Adaptation conditions. This insensitivity to indexical information in the context of sensitivity to phonetic change suggests that the anterior portion of the mid-STG bilaterally is speaker invariant.

The locus of this invariance effect differs from that shown for phonetic category invariance which emerged in the left IFG (Myers et al., 2009) suggesting that the neural system treats these two types of variability separately. That speaker insensitivity to phonetic change emerged bilaterally in the temporal lobe suggests that this computation occurs early in the processing stream. Consistent with this claim is behavioral evidence showing continuous perception for a male-female synthetic [i] continuum in contrast to the typical categorical perception function found for phonetic category continua varying along a particular phonetic dimension (Mullenix, Johnson, Topcu-Durgan, & Farnsworth, 1995; see also Mullinex, 1997 for a review). The shape of the function for the perception of the speaker continuum was similar to that obtained with low-level auditory stimuli such as tones, suggesting that the resolution of voice information occurs at an early, auditory level of processing.

Myers et al. (2009) proposed that phonetic category invariance arose from decision-related mechanisms in the prefrontal cortex and that this area plays a role in computing category representations. The emergence of speaker invariance in the STG suggests that the computations over which speaker invariance is derived are based instead on sensory properties of the input rather than on decision-related properties. Of interest, Myers et al. showed functional heterogeneity in the IFG with different parts of the IFG showing sensitivity to phonetic category change and others showing phonetic category invariance. Similarly, the STG appears to show functional heterogeneity with the posterior STG failing

to show speaker invariance but instead showing increased activation to phonetic change when spoken by different speakers compared to phonetic change spoken by the same speaker. Additionally, a recent study by Leaver and Rauschecker (2010) compared activation of two conditions in which either phonetic or speaker identity changed, and reported a region in the left mid anterior STG that showed greater activation to changing phonetics. The coordinates reported for this region (X, Y, Z = −60, −20, 1) fall within our left anterior mid STG cluster (X,Y,Z = −61, −19, 2).

Functional Architecture of Speech Processing

This is the first study to our knowledge that demonstrates true speaker invariance – that is sensitivity to phonetic information and insensitivity to speaker information. The use of a condition in which change occurs across both phonetic and speaker dimensions allowed for the determination of speaker insensitivity by showing that there is not merely more activation for phonetic processing than speaker processing in these regions, but there is no change in activation with the addition of speaker variability. The regions showing this pattern are located in the anterior STG bilaterally. That these regions are so close to the primary auditory cortex suggests that 1) speaker variability is treated differently than phonetic variability and 2) the superior temporal cortex is involved in early categorization of auditory objects.

The evidence for speaker invariance provides support for the view that speaker normalization processes allow for the translation of a variable speech input to be mapped on to a common abstract sound structure. That this process appears to occur early in the processing stream, recruiting temporal structures, suggests that this mapping takes place prelexically, before sound structure input is mapped on to lexical representations. Recent behavioral evidence is consistent with this view. McQueen et al. (2006) showed that subjects can be trained to retune their perception of the boundaries between phonetic categories (in their case, word final [f] and [s]) given a particular training set of words, and this retuning generalized across words outside of the training set. Thus, the reshaping of phonetic categories is not limited to specific ‘episodes’, i.e. instances of the heard words, but extends to other words in the lexicon (see McQueen et al. 2006 for discussion).

Our finding of regions insensitive to speaker information necessitates a reconciliation between the large body of behavioral evidence supporting the episodic theory of speech perception that posits retained indexical information on phonetic and lexical representations. One possible interpretation is that multiple processing streams are in play, one of which extracts phonetic information from speech input, and another in which the input is processed holistically. In visual processing, evidence for a dual processing stream comes from Konen & Kastner (2008), who proposes two hierarchical pathways for visual object recognition in the temporal and parietal lobes. It is not surprising that multiple systems would be involved in visual object recognition given the many sources of variability (viewing angle, distance, rotation, lighting etc.). Likewise, it is likely that multiple streams would be required for the numerous sources of variability in auditory object recognition (phonetic context, phonetic variability, speaker identity, affect, environmental noise, etc.). The presence of an array of mechanisms for achieving invariance for different sources of variability, as well as a holistic process in which word recognition happens without normalization would offer built in redundancy required for the difficult task of speech perception.

The current findings do not rule out the possibility that input to the lexicon contains both abstract representations and detailed acoustic episodes. Indeed, it is clear from behavioral studies that listeners are sensitive to and their performance is influenced by talker variability (Goldinger, 1998). Moreover, individual attributes of a talker’s voice allows for knowing which phonetic attributes are idiosyncratic and which ones have phonetic significance

(Eisner and McQueen, 2005). What the results of the current study do suggest is that at some level of processing the variable speech input is mapped on to abstract properties of phonetic categories and lexical form.

Acknowledgments

This research was supported in part by NIH NIDCD Grant RO1 DC006220 to Brown University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Deafness and Other Communication Disorders or the National Institutes of Health. Many thanks to John Mertus for his help in programming the experiment and hardware support and to Kathy Kurowski for her assistance in the recording the stimuli.

References

Belin P, Zatorre RJ. Adaptation to speaker's voice in right anterior temporal lobe. Neuro Report. 2003; 14:2105– 2109.

Belin P, Fecteau S, Bedard C. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences. 2004; 8:129–135. [PubMed: 15301753]

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000; 403:309– 312. [PubMed: 10659849]

Binder, JR.; Price, C. Functional neuroimaging of language. In: Cabeza, R.; Kingstone, A., editors. Handbook of Functional Neuroimaging of Cognition. Cambridge: MIT Press; 2001. p. 187-251.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET. Human temporal lobe activation by speech and non-speech sounds. Cerebral Cortex. 2000; 10:512–528. [PubMed: 10847601]

Blumstein SE, Myers EB, Rissman J. The Perception of Voice Onset Time: An fMRI Investigation of Phonetic Category Structure. Journal of Cognitive Neuroscience. 2005; 17:1353–1366. [PubMed: 16197689]

Boersma P. Praat, a system for doing phonetics by computer. Glot International. 2001; 5:341–345.

Britton B, Blumstein SE, Myers EB, Grindrod C. The role of spectral and durational properties on hemispheric asymmetries in vowel perception. Neuropsychologia. 2009; 47:1096–1106. [PubMed: 19162052]

Celsis P, Boulanouar K, Doyon B, Ranjeva JP. Differential fMRI responses in left superior temporal gyrus and left supramarginal gyrus to habituation and change detection in syllables and tones. Neuro Image. 1999; 9:133–144.

Cox RW. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. Computers and Biomedical Research. 1996; 29:162–173. [PubMed: 8812068]

Davis MH, Johnsrude IS. Hierarchical processing in spoken language comprehension. Journal of Neuroscience. 2003; 23:3423–3431. [PubMed: 12716950]

Eisner F, McQueen JM. The specificity of perceptual learning in speech processing. Perception and Psychophysics. 2005; 67:224–238. [PubMed: 15971687]

Fecteau S, Armony JL, Joanette Y, Belin P. Is voice processing species specific in human auditory cortex? An fMRI study. Neuroimage. 2004; 23:840–848. [PubMed: 15528084]

Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 2008; 322:970–973. [PubMed: 18988858]

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. Visual categorization and the primate prefrontal cortex: neurophysiology and behavior. J Neurophysiol. 2002; 88:929–941. [PubMed: 12163542]

Freedman DJ, Riesenhuber M, Poggio T, Miller EK. A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci. 2003; 23:5235–5246. [PubMed: 12832548]

Giraud AL, Kell C, Thierfelder C, Sterzer P, Russ MO, Preibisch C. Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. Cerebral Cortex. 2004; 14:247–255. [PubMed: 14754865]

Goldinger SD. Echoes of echoes?: An episodic theory o lexical access. Psychological Review. 1998; 105:251–279. [PubMed: 9577239]

Grill-Spector K, Malach R. fMR-adaptation: a tool for studying the functional properties of human cortical neurons. Acta Psychologica. 2001; 107:293–321. [PubMed: 11388140]

Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews Neuroscience. 2007; 8:393–402.

Jiang X, Bradley E, Rini R, Zeffiro T, VanMeter J, Riesenhuber M. Categorization training results in shape and category-selective human neural plasticity. Neuron. 2007; 53:891–903. [PubMed: 17359923]

Joanisse MF, Zevin JD, McCandliss BD. Brain mechanisms implicated in the preattentive categorization of speech sounds revealed using fMRI and a short-interval habituation trial paradigm. Cerebral Cortex. 2007; 17:2084–2093. [PubMed: 17138597]

Konen CS, Kastner S. Two hierarchically organized neural systems for object information in human visual cortex. Nature Neuroscience. 2008; 11:224–231.

Leaver AM, Rauschecker JP. Cortical Representation of Natural Complex Sounds: Effects of Acoustic Features and Auditory Object Category. The Journal of Neuroscience. 2010; 30(9):7604–7612. [PubMed: 20519535]

Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA. Neural substrates of phonemic perception. Cerebral Cortex. 2005; 15:1621–1631. [PubMed: 15703256]

McQueen JM, Cutler A, Norris D. Phonological abstraction in the mental lexicon. Cognitive Science. 2006; 30:1113–1126. [PubMed: 21702849]

Mertus, JA. Brown Lab Interactive Speech System; 2009. http://www.metus.org/Bliss

Mullenix, JW. On the nature of perceptual adjustments to voice. In: Johnson, K.; Mullenix, JW., editors. Talker variability in speech processing. New York: Academic Press; 1997.

Mullenix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception and Psychophysics. 1990; 47:379–390. [PubMed: 2345691]

Mullenix JW, Johnson KA, Topcu-Durgan M, Farnsworth LM. The perceptual representation of voice gender. Journal of the Acoustical Society of America. 1995; 98:3080–3095. [PubMed: 8550934]

Myers EB, Blumstein SE, Walsh E, Eliassen J. Inferior frontal regions underlie the perception of phonetic category invariance. Psychological Science. 2009; 20:895–903. [PubMed: 19515116]

Nearey TM. Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America. 1989; 85:2088–2113. [PubMed: 2659638]

Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971; 9:97–113. [PubMed: 5146491]

Peretz I, Kolinsky R, Tramo M, Labrecque R, Hublet C, Demeurisse G. Functional dissociations following bilateral lesions of auditory cortex. Brain. 1994; 117:1283–1301. [PubMed: 7820566]

Scott SK, Wise RJS. The functional neuroanatomy of prelexical processing in speech perception. Cognition. 2004; 92:13–45. [PubMed: 15037125]

Scott SK, Blank CC, Rosen S, Wise RJ. Identification of a pathway for intelligible speech in the left temporal lobe. Brain. 2000; 123:2400–2406. [PubMed: 11099443]

Stevens KN. Toward a model for speech recognition. Journal of the Acoustical Society of America. 1960; 32:47–55.

Van Lancker DR, Canter GJ. Impairment of voice and face recognition in patients with hemispheric damage. Brain Cognition. 1982; 1:185–951.

Van Lancker DR, Kreiman J, Cummings J. Voice perception deficits: Neuroanatomical Correlates of Phonagnosia. Journal of Clinical and Experimental Neuropsychology. 1989; 11:665–674. [PubMed: 2808656]

von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Research Cognitive Brain Research. 2003; 17:48–55. [PubMed: 12763191]

von Kriegstein K, Smith DR, Patterson RD, Ives DT, Griffiths TD. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Current Biology. 2007; 17:1123–1128. [PubMed: 17600716]

von Kriegstein K, Warren JD, Ives DT, Patterson RD, Griffiths TD. Processing the acoustic effect of size in speech sounds. Neuroimage. 2006; 32:368–375. [PubMed: 16644240]

von Kriegstein K, Smith DR, Patterson RD, Kiebel SJ, Griffiths TD. How the Human Brain Recognizes Speech in the Context of Changing Speakers. The Journal of Neuroscience. 2010; 30(2):629–638. [PubMed: 20071527]

Warren JD, Scott SK, Price CJ, Griffiths TD. Human brain mechanisms for the early analysis of voices. Neuro Image. 2006; 31:1389–1397. [PubMed: 16540351]

Wong PCM, Nusbaum HC, Small SL. Neural bases of talker normalization. Journal of Cognitive Neuroscience. 2004; 16:1173–1184. [PubMed: 15453972]

Zevin JD, Yang J, Skipper JI, McCandliss BD. Domain General Change Detection Accounts for "Dishabituation" Effects in Temporal–Parietal Regions in Functional Magnetic Resonance Imaging Studies of Speech Perception. The Journal of Neuroscience. 2010; 30(3):1110–1117. [PubMed: 20089919]

Zevin JD, McCandliss BD. Dishabituation of the BOLD response to speech sounds. Behavioral and Brain Functions. 2005; 1:4. [PubMed: 15953396]

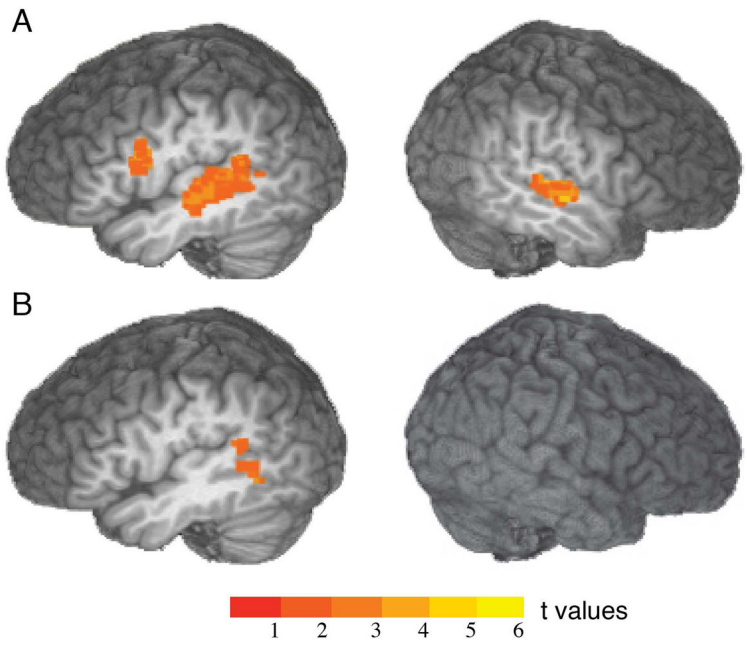

Figure 1.
Clusters significant at a corrected threshold of $p<0.05$ ($p<0.025$ voxel-wise threshold, minimum cluster size 33 voxels) for Phonetic Change versus Adaptation (A) and Speaker Change versus Adaptation (B).

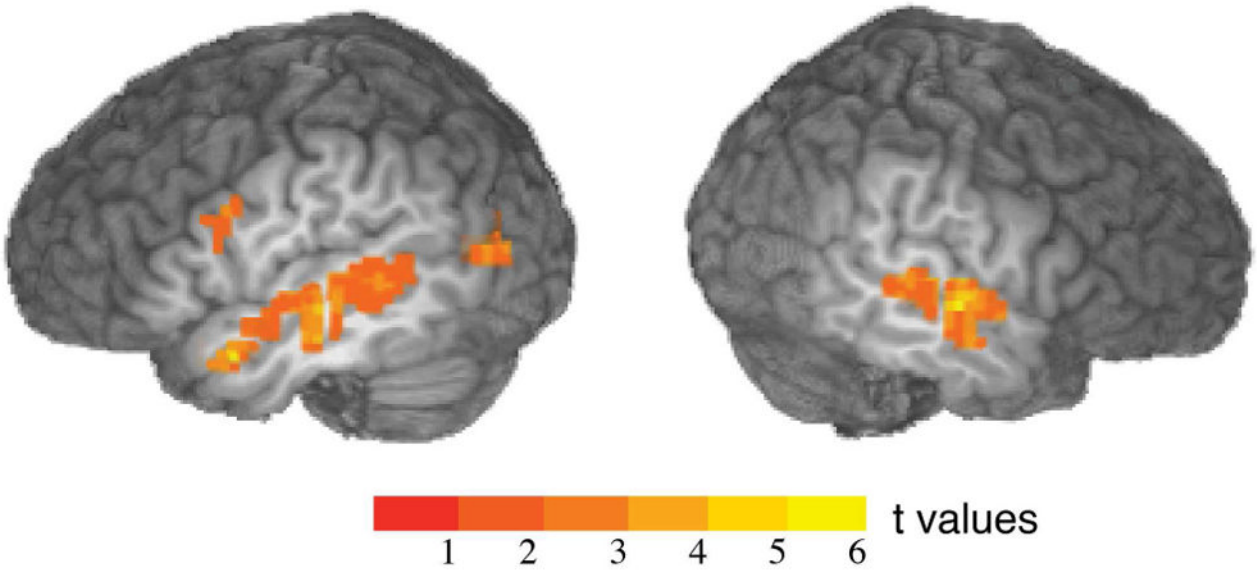

Figure 2.
Clusters significant at a corrected threshold of p<0.05 (p<0.025 voxel-wise threshold, minimum cluster size 33 voxels) for Phonetic Change & Both Change versus Speaker Change & Adaptation.

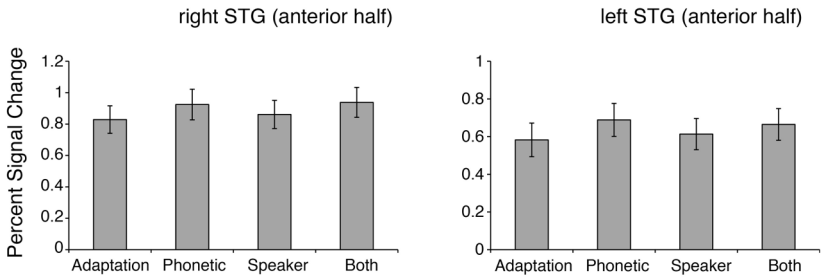

Figure 3.
Region of interest analysis showing percent signal change across conditions for the two clusters that demonstrated the speaker invariance pattern.

Table 1

Accuracy and reaction-time latencies (RT) for different responses on the behavioral pretest.

Condition Task	Speaker Change Speaker Task	Phonetic Change Phonetic Task	Both Change Speaker Task	Both Change Phonetic Task
% correct	99	99	100	99
RT (SE)	803 (36)	851 (42)	752 (38)	823 (51)

Table 2

Summary of activated areas. Coordinates indicate the location of the maximum t-stats within the cluster.

Anatomical Region	Voxels	Max t-stat	P value	x	y	z
Phonetic Change vs Adaptation						
LSTG	249	4.366	p<0.0001	−62	−26	12
RSTG	93	5.221	p<0.0001	59	−17	9
Left IFG (BA 44)	46	4.283	p<0.0064	−44	5	24
Speaker Change vs Adaptation						
LSTG	48	4.012	p<0.0048	−40	−49	5
Phonetic Change & Both vs Speaker Change & Adaptation						
LSTG (posterior half)	132	4.237	p<0.0001	−61	−25	5
L post MTG	106	4.757	p<0.0001	−46	−70	17
RSTG (posterior half)	97	5.643	p<0.0001	49	−22	8
L temporal pole	77	5.136	p<0.0002	−49	4	−15
RSTG (anterior half)	72	6.090	p<0.0002	55	−13	2
LSTG (anterior half)	68	4.420	p<0.0004	−61	−19	2
LIFG (opercularis)	33	3.84	p<0.0438	−37	10	29