Available online at www.sciencedirect.com
ScienceDirect
Brain and Language 98 (2006) 310–318
Brain and Language
www.elsevier.com/locate/b&l

Training to use voice onset time as a cue to talker identification induces a left-ear/right-hemisphere processing advantage ☆

Alexander L. Francis *, Courtney Driscoll

Department of Speech, Language and Hearing Sciences, Purdue University, 1353 Heavilon Hall, 500 Oval Drive, West Lafayette, IN 47907, USA

Accepted 1 June 2006
Available online 7 July 2006

Abstract

We examined the effect of perceptual training on a well-established hemispheric asymmetry in speech processing. Eighteen listeners were trained to use a within-category difference in voice onset time (VOT) to cue talker identity. Successful learners ($n = 8$) showed faster response times for stimuli presented only to the left ear than for those presented only to the right. The development of a left-ear/right-hemisphere advantage for processing a prototypically phonetic cue supports a model of speech perception in which lateralization is driven by functional demands (talker identification vs. phonetic categorization) rather than by acoustic stimulus properties alone.

Keywords: Hemispheric asymmetry; Talker identification; Voice onset time; VOT; Functional lateralization

1. Introduction

Neuropsychological and neuroimaging research suggests that speech processing is asymmetrically distributed between the two cerebral hemispheres. The left hemisphere, especially the mid- and posterior-superior temporal gyrus (STG) and superior temporal sulcus (STS) is preferentially involved in linguistic (specifically phonological) processing (Giraud & Price, 2001; Indefrey & Cutler, 2005; Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Scott, Blank, Rosen, & Wise, 2000). By contrast, the right hemisphere, especially anterior STS, is preferentially involved in processing human voices (Belin & Zatorre, 2003; Belin, Zatorre, & Ahad, 2002; Imaizumi et al., 1997; Levy, Granot, & Bentin, 2001, 2003; Nakamura et al., 2001; Van Lancker, Kreiman, & Cummings, 1989; Von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003; Von Kriegstein & Giraud, 2004). One explanation for this distribution is that the two hemispheres are specialized for processing different kinds of acoustic events. According to one such model, the left hemisphere is specialized for processing rapidly changing temporal events, while the right is specialized for processing spectral properties (Zatorre & Belin, 2001; Zatorre, Belin, & Penhune, 2002). In contrast, it has also been proposed that the hemispheres differ primarily in terms of the temporal width of their "window of analysis," with the left hemisphere preferentially processing stimuli that change within a window of 20–50 ms, and the right focusing on longer-term changes on the order of 150–300 ms (Boemio, Fromm, Braun, & Poeppel, 2003; Poeppel, 2003). Both of these models, however, make similar predictions: The left hemisphere should be superior at processing rapidly changing acoustic properties such as formant transitions and voice onset time differences that typically cue phonological contrasts and occur within a few tens of milliseconds, while the right hemisphere should excel at processing more gradually changing properties such as

☆ Results presented here derive from work submitted by the second author as part of a thesis in partial fulfillment of the requirements for the degree of Master of Science in Speech and Hearing Sciences at Purdue University. We are grateful to Jack Gandour, Natalya Kaganovich, David Kemmerer, Robert Melara, Christine Weber-Fox and two anonymous reviewers for helpful suggestions and discussion related to earlier versions of this article. A preliminary version of these results was presented at the 2005 ASHA Convention, San Diego, CA, November 19, 2005.
* Corresponding author. Fax: +1 765 494 0771.
E-mail address: francisa@purdue.edu (A.L. Francis).

0093-934X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.bandl.2006.06.002

fundamental frequency and long-term average spectrum that are thought to play a role in talker identification and unfold over hundreds of milliseconds.

One possible shortcoming of this hypothesis is that, in most experiments, the acoustic properties of the stimuli are typically confounded with task: studies of phonological processing typically employ segmental stimuli (e.g., consonant–vowel (CV) syllables) that differ primarily according to rapidly changing acoustic properties, while studies of voice processing typically employ much longer stimuli (e.g., sentences). This confound is not easily avoided. Most of the acoustic properties associated with phonetic categorization are in fact relatively brief (on the order of tens to hundreds of milliseconds), and typically involve rapid spectral changes (Raphael, 2005). In contrast, those associated with talker identity typically involve much longer-term variability, and often unfold completely only over the course of much longer utterances (Kreiman, Van Lancker-Sidtis, & Gerratt, 2005).

Recent research by Gandour and colleagues avoids this confound by investigating the perception of lexical tones, and supports the hypothesis that hemispheric processing of linguistic information depends at least partly on functional rather than purely acoustic factors (see Gandour & Dzemidzic et al., 2003 for discussion). For example, Gandour et al. (2004) compared brain activation differences in two groups of listeners (Thai and Chinese) discriminating pairs of Thai tonal contours presented in speech (natural syllables) or non-speech. In the speech condition, Thai speakers showed left-hemisphere fronto-parietal activation, suggesting that they were processing the information phonologically. In contrast, Chinese speakers did not show this activation pattern, suggesting that they were discriminating sounds according to non-linguistic criteria. Similarly, Gandour et al. (2002) found that Chinese speakers show heightened left-hemisphere activation for processing tonal contours as lexical tones, but increased right-hemisphere activation when processing these same contours as changes in intonation. Although the acoustic properties of the tonal contours were the same across conditions, processing was lateralized differently depending on the listeners' functional goal, demonstrating flexibility in the lateralization of function independent of acoustic feature (i.e., using pitch as a cue to intonational vs. lexical phonology).

The present paper examines a similar functional dependency in lateralization, but in this case training was used to shift processing of a typically left-hemispheric acoustic cue to the right. Our goal was to distinguish between a stimulus property-based account of lateralization and one based on the functional goal of the task. Listeners were trained to identify stimuli that had been produced by a single male talker as having been produced by two different talkers. Original recordings of one person saying voiceless, aspirated consonant–vowel (CV) syllables were resynthesized to create pairs of tokens that differed only in voice onset time (VOT). VOT is a rapidly changing temporal property associated in English with the distinction between voiced and voiceless stop consonants (/b/, /d/, and /g/ vs. /p/, /t/, and /k/). Thus, an acoustically based account of lateralization would predict that it should be processed preferentially in the left hemisphere.[1] On the other hand, since identical stimuli were used before and after training, any change in hemispheric dominance for processing this quintessentially left-hemisphere acoustic property would provide strong support for a model of lateralization based on function, not (only) stimulus properties.

If the right-hemisphere preference for talker-related processing is based on the acoustic properties of talker-related cues (e.g., pitch, long-term average spectrum), then training listeners to identify our stimuli as having been produced by two different talkers should not affect lateralization of processing because the stimuli do not differ according to properties typically associated with talker identification. However, if talker identification-training can induce an increased left-ear/right-hemisphere advantage for processing of VOT, this would indicate that the right-hemisphere preference for processing talker-related information is based on the task (talker identification) rather than the acoustic properties being judged (VOT). That is, an increased left-ear/right-hemisphere advantage would suggest that the right hemisphere is specifically adapted to processing talker-related information regardless of the acoustic properties that encode it.

2. Methods

2.1. Subjects

Eighteen participants (8 men, 10 women) between the ages of 18 and 35 successfully completed the experiment after providing informed consent. All participants were native speakers of a North American dialect of English and had no significant foreign language experience. Participants reported no history of neurological disorder. All demonstrated hearing within normal limits bilaterally as determined by pure tones presented at 25 dB HL at 500, 1000, 2000, and 4000 Hz through a GSI-61 portable audiometer. All were right-hand dominant as determined by a standard handedness questionnaire (Oldfield, 1971). Twelve additional participants (4 men, 8 women) meeting the same criteria were recruited to serve as a control group and were trained to identify the same stimuli as members of different

[1] Note that Segalowitz and Cohen (1989) reports a right-hemisphere advantage for processing VOT differences in a non-response condition, and argue that the LH dominance for speech (especially consonant) processing may result from an attentional bias toward LH processing arising from subvocal articulation in conditions requiring a response (even a non-vocal one). In the present experiment, however, listeners *always* responded and therefore were expected to show the typical LH bias for processing VOT (e.g., Schwartz & Tallal, 1980) although this bias was not in fact identified (see Footnote 7, below).

phonetic categories rather than as the productions of two different talkers.

2.2. Stimuli

Speech tokens were derived from a natural male voice saying consonant–vowel (CV) syllables. Syllables recorded were all English voiceless aspirated stops followed by three cardinal vowels (high front, high back, and low central): [pa], [pi], [pu], [ta], [ti], [tu], [ka], [ki], and [ku] (one recording of each syllable was used). Using Praat 4.2 (Boersma & Weenink, 2006), stimuli were digitized at 22.05 kHz with 16-bit quantization, normalized in peak amplitude, and the duration of each stimulus was adjusted to between 260 and 280 ms (slight differences across vowels and consonant place of articulation were maintained to reflect natural variability). The syllables were then duplicated and tokens with long (50 ms) and short (30 ms) voice onset time (VOT) were created using PSOLA resynthesis to either expand or compress the existing VOT period for each stimulus and then adjusting the duration of the vowel in the opposite direction to maintain overall syllable duration. Exact VOT durations differed slightly across place of articulation by as much as 3 ms. Long-VOT tokens were arbitrarily defined as having been produced by "Dave" while short-VOT tokens were defined as having been produced by "Jared." Listeners were not told that the stimuli were derived from productions of a single talker until the experiment was complete.

2.3. Procedure

The experiment consisted of three stages (pre-training test, training, and post-training test), each lasting approximately 1 h on three different days over a 1-week time period. In each stage, all stimuli were presented at a comfortable listening level through a pair of Sennheiser HD-25 headphones at a sampling rate of 22.05 kHz with 16-bit quantization. All tests and training sessions were conducted using E-Prime 1.1 (Schneider, Eschman, & Zuccolotto, 2002a, 2002b, 2002c) combined with a Cedrus RB-620 six-button response box recording accuracy (proportion correct) and response time with millisecond accuracy.

On each test trial, participants heard a stimulus and made a response by pressing one of two buttons on a response box. The two response choices ("Dave" and "Jared") were displayed simultaneously with auditory presentation of the stimulus. Response time for each trial was limited to 5 s, but the task was otherwise self-paced. Listeners were asked to respond accurately, and were informed that there was a 5 s upper limit on response times. Responses longer than 5 s were counted as incorrect. No auditory or visual feedback was provided during testing.

For each of the pre-training test and post-training test sessions, a total of 576 tokens of both long-VOT and short-VOT versions of all stimuli were presented *monaurally*.[2] Presentations were randomized by ear and VOT but were blocked by place of articulation and left–right response order. Left–right order of responses was counterbalanced within subjects (e.g., in blocks where "Dave" was shown on the left, pressing the left button indicated that "Dave" had produced the syllable). Thus, each of nine tokens was presented 16 times to each ear, eight times with "Dave" on the left and eight times with "Dave" on the right. Thus, three different places of articulation $\times$ 3 vowels $\times$ 2 VOT-lengths $\times$ 2 ears $\times$ 2 orders of presentation $\times$ 8 repetitions = 576 tokens.[3]

Training trials were identical to test trials except that trial-level visual and auditory feedback was provided. Immediately after each response participants either saw the word "Correct!" or "Incorrect" along with the correct talker's name on the screen followed by an auditory repetition of the stimulus 750 ms after the onset of visual feedback. Visual feedback remained on the screen for 1500 ms. A lack of response was treated as an error except that "No response detected" appeared on the screen instead of "Incorrect."

In training, 640 tokens were presented *binaurally*. The training set was a subset of the test set, consisting of both Dave (long-VOT) and Jared (short-VOT) versions of [ki], [ku], [ti], and [tu]. Thus, all [p] and all [a] syllables were heard only in testing. Participants heard each token 80 times (two places of articulation $\times$ 2 vowels $\times$ 2 VOT-lengths $\times$ 80 trials = 640 tokens).

Immediately preceding the post-training test, each participant completed a "mini-training" program identical in structure to the larger training program but using only 32 trials (four presentations of each training token) to reinforce the effects of the relatively short training period

[2] A monaural testing procedure was used for a variety of reasons. First, a behavioral paradigm was chosen because it was simpler to carry out and less resource-intensive than the kinds of imaging methods (PET, FMRI, MEG and ERP) that are now commonly used for investigations of hemispheric asymmetry. Both monaural and dichotic presentation have been used successfully to investigate hemispheric asymmetries in auditory processing, and previous research has showed that results attained by the two tasks are comparable (Bever, 1971; Bradshaw & Nettleton, 1988; Friedrich, 1974). Dichotic listening requires presentation of two conflicting stimuli to separate ears (e.g., two different talkers), but in the present experiment the two "different" talkers were necessarily so similar that it seemed possible that dichotic presentation would reinforce the perception that the two stimuli were produced by a single talker, thereby undermining any effect of training. Thus, monaural presentation provided the optimal combination of ease of data collection, and high probability of obtaining interpretable results.

[3] Note that Rimol, Eichele, and Hugdahl (2006) showed that long VOT (voiceless aspirated) Norweigian stimuli tend to be perceived more easily than short VOT (voiceless unaspirated) tokens, and that this difference led to a processing advantage for the ear to which the long-VOT token was presented. In the present study, long- and short-VOT tokens were presented an equal number of times to each ear, RTs were averaged over correct responses irrespective of VOT duration and all VOT values for both short- and long-VOT tokens were well within the boundaries of the voiceless aspirated category of English. Moreover, a two-way, repeated measures ANOVA of proportion correct with factors of test (pre-training and post-training) and Length (short-VOT and long-VOT) showed the expected effect of test, $F(1,7) = 5.88$, $p = .05$, but no effect of Length, $F(1,7) = 3.31$, $p = 0.1$, and no interaction, $F(1,7) = 1.90$, $p = .21$.

(cf. Francis, Baldwin, & Nusbaum, 2000). The post-training test was identical in structure to the pre-training test.

Participants in the control (phonetic training) group were trained and tested using identical methods, except that their response choices were based on a phonetic voicing contrast rather than talker identity. Tokens with a long VOT ("Dave") were treated as voiceless ([p], [t], or [k]) while those with a short VOT ("Jared") were treated as voiced ([b], [d], or [g]). Since trials were blocked by place of articulation, on any given trial listeners in this group also chose between two possible responses (e.g. [p] vs. [b] instead of "Dave" vs. "Jared"). All other aspects of training and testing were identical across the two groups.

2.4. Analysis

The focus of this experiment is on changes in ear advantage related to successful learning. Therefore, primary analyses were conducted only on data from eight individuals who improved as a result of training by at least five percentage points on identification of all test stimuli (including both generalization and training tokens). Similarly, because listeners have been shown to generally exhibit the largest effect of training when tested on the same tokens they were trained on (e.g., Francis et al., 2000; see Logan & Pruitt, 1995 for discussion), only responses to stimuli in the training set were analyzed completely. However, comparisons are made between learners and non-learners, and between results for training-set and generalization stimuli when appropriate. In all cases, prior to statistical analysis, all proportions were arcsine transformed and all response times were logarithmically transformed (Kirk, 1995).[4] When considering response times, measurements were made from stimulus onset, only correct responses were analyzed, and, following the discussion of Ratcliff (1993), outliers (responses greater than 2.5 standard deviations above the mean for each subject across all conditions) were excluded (approximately 4% of all correct responses).

3. Results

On debriefing, participants reported that, at first, the task was very difficult but by the end of training they were confident that there were in fact two different talkers. Response times (RTs) to correct identifications were analyzed using a two-way, repeated measures ANOVA with two levels of test (pre-training and post-training) and two levels of ear (left and right) that showed no significant effect of test, $F(1,7) = 0.176$, $p = .69$, or of ear, $F(1,7) = 1.32$, $p = .29$, but did show a significant interaction of test by ear, $F(1,7) = 6.77$, $p = .04$ (shown in Fig. 1). Post hoc (Tukey HSD) analysis showed that this interaction reflected a significant ($p = .05$) difference on the post-training test between correct responses to stimuli presented to the left-ear (1165 ms) as opposed to the right (1257 ms), although there was no difference between ears on the pre-training test (1229 ms vs. 1225 ms, respectively). This pattern is consistent with what we would expect if the relative dominance of the right hemisphere were increasing, leading to faster processing of stimuli presented to the left-ear (right hemisphere) and/or slower processing of stimuli presented to the right-ear (left hemisphere).

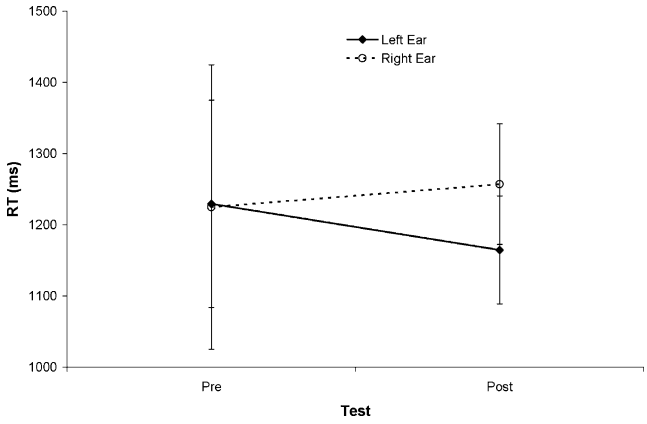

Fig. 1. Response times for correct responses on pre-training test and post-training test, by ear. Error bars indicate standard error of the mean.

3.1. Generalization

Training was very brief (around 3 h) in order to allow the study to be carried out in a reasonable amount of time per subject, and therefore learning was limited in scope.[5] In particular, listeners showed only a small degree of generalization of learning. A two-way, repeated measures ANOVA comparing two levels of token set (training and generalization) and two levels of test (pre-training and post-training) showed the expected significant effect of test, $F(1,7) = 17.03$, $p = .004$, as learners improved overall from 51% to 61% correct, and a significant effect of set, $F(1,7) = 13.26$, $p = .008$. There was no interaction between the two, $F(1,7) = 4.08$, $p = .08$, but post hoc (Tukey HSD) analysis showed that the improvement in training set tokens from 53% to 70% correct was significant ($p = .03$) while the change in generalization set tokens from 48% to 53% correct was not ($p = .08$). The main effect of set and the lack of an interaction might suggest that training set tokens were inherently easier to identify than generalization set tokens. However, post hoc analysis showed that this effect was due primarily to the large proportion of correct responses to the training set stimuli following training:

[4] Following the suggestions of Ratcliff (1993), analyses were repeated using the inverse transformation, providing comparable results.

[5] Note that Pisoni, Aslin, Perey, and Hennessy (1982) found 2 h sufficient to induce listeners to perceive a new phonetic category defined by VOT alone. Thus, it was expected that 3 h should be sufficient in the present case as well. However, subjects in the Pisoni et al. (1982) experiment were asked to learn a different category boundary location, one that may be easier to learn than that used in the present experiment due to the location of regions of heightened psychoacoustic sensitivity (cf. discussion by Holt, Lotto, & Diehl, 2004).

there was no significant difference between the two types of stimuli on the pretest ($p = .77$) but there was on the post-test ($p = .03$). Thus, after training, training set tokens were identified better than generalization set tokens, but it is not clear from the accuracy data alone whether this effect is due to training alone, or whether the training set tokens were for some reason easier to learn than the generalization set tokens.

An analysis of response times, however, suggested that, before training, training set tokens were more difficult to identify than generalization set tokens, but after training there was no significant difference between the two. A two-way repeated measures ANOVA with two levels of test (pre-training and post-training) and two levels of set (training and generalization) and subsequent post hoc (Tukey HSD) analyses showed no main effect of test, $F(1,7) = 1.46$, $p = .27$, or of Set, $F(1,7) = 0.93$, $p = .37$, but there was a significant interaction between the two, $F(1,7) = 18.41$, $p = .004$, reflecting a significant difference in response times on the pretest between training set tokens (1227 ms) and generalization set tokens (1112 ms) but no difference between the two on the post-test (1211 ms vs. 1303 ms).

Analyzing only the generalization tokens using a repeated measures ANOVA with two levels of test (pre-training and post-training) and two levels of ear (left and right) showed that learners actually *increased* the time it took to make a correct response to tokens in the generalization set from 1112 to 1303 ms, $F(1,7) = 5.64$, $p = .05$, but there was no effect of ear, $F(1,7) = 0.02$, $p = .91$, and no interaction, $F(1,7) = 0.02$, $p = .90$.

Generalization tokens also differed from one another in terms of how quickly listeners were able to respond correctly to them. A distinction can be made between generalization tokens that differed from training set tokens only according to the vowel ([ta] and [ka], the "vowel" tokens), those that differed only according to consonant ([pi] and [pu], the "consonant" tokens) and those that differed according to both consonant and vowel ([pa], the "both" tokens). A two-way repeated measures analysis of variance with two levels of test (pre-training and post-training) and three levels of Type (vowel, consonant, both) showed a significant effect of Type, $F(2,14) = 4.98$, $p = .02$, but no effect of test, $F(1,7) = 5.15$, $p = .06$, and no interaction, $F(2,14) = 0.94$, $p = .42$. Post hoc (Tukey HSD) analysis showed a significant difference only between "vowel" and "both" tokens ($p = .03$). Tokens differing from the training set only in vowel were identified most slowly (1237 ms), while those differing only in consonant were identified slightly (not significantly) more quickly (1211 ms), and those differing in both were identified most quickly (1124 ms). More importantly, post hoc analysis of the (non-significant) interaction revealed still more differences between these three types of generalization tokens. "Vowel" tokens showed no significant increase in RT between pre-training (1177 ms) and post-training (1298 ms), but "consonant" and "both" tokens did, with "consonant" tokens increasing from 1095 to 1327 ms (232 ms) and "both" tokens increasing from 1005 to 1243 ms (238 ms). Thus, although the [pa] tokens were overall significantly easier to identify than the [ti] or [tu] tokens, training caused listeners to take significantly more time on both the [pa] and the [pi] and [pu] tokens but not on the [ta] and [ka] tokens. These results suggest that learning was specific to the consonants [t] and [k], and are consistent with recent results presented by Eisner and McQueen (2005) showing that short-term perceptual learning can be specific to particular phonemes.

3.2. Non-learners

In the absence of additional evidence, it might be argued that the observed pattern of change in hemispheric dominance for processing VOT is simply the result of exposure to stimuli produced by a particular voice. Belin and Zatorre (2003) showed that merely adapting to a talker's voice can in itself induce appreciable changes in right-hemisphere activation, and it is important to rule out the possibility that something similar might be happening here. If the present results were an effect of adaptation to these specific stimuli, then all listeners exposed to these stimuli should have shown a similar pattern of change in hemispheric dominance, but this was not the case. We also analyzed the performance of the 10 listeners who were exposed to these stimuli under exactly the same conditions but showed no appreciable learning. Each of these non-learners showed an improvement of less than 5 percentage points from pre-training test to post-training (mean = 0.49% improvement, ranging from 4.7% improvement to −2.8% decrement). A repeated measures ANOVA of response times to correct responses with two levels of test (pre-training and post-training) and two levels of ear (left and right) showed no significant effects of test, $F(1,9) = 2.28$, $p = .17$, or of ear, $F(1,9) = 0.202$, $p = .66$, and no interaction between the two, $F(1,9) = 4.10$, $p = .07$. Thus, any change in hemispheric processing exhibited by the successful learners must have resulted from successful learning, not mere exposure to stimuli that they were told were produced by two different voices.

Learners and non-learners also differed in terms of their treatment of generalization tokens, as shown by the results of a mixed factorial ANOVA with two levels of Group (learners and non-learners), two levels of test (pre-training and post-training) and two levels of token Type (training and generalization) with subsequent post hoc (Tukey HSD) analyses. While the learners showed a non-significant 16 ms decrease in RT for trained tokens from 1227 to 1211 ms and a significant 191 ms increase in RT for generalization tokens from 1112 to 1303 ms, non-learners showed similar but non-significant increases in RT for both the training tokens (64 ms, from 979 to 1043 ms) and the generalization tokens (88 ms, from 906 to 994 ms). This pattern of increasing response time suggests that successful learning somehow increased processing demands for generalization-set stimuli.

3.3. Phonetic training

It is also possible that the observed change in hemispheric dominance reflects the effect of improved categorization, regardless of the kind of categorization being learned. However, this hypothesis was contradicted by the results of a study of eight listeners[6] who were successfully trained to hear these stimuli as members of different phonetic classes. These participants were trained using the identical training methods and number of trials as those in the main study, except that these listeners were trained to label the short-VOT ("Jared") stimuli as [b], [d], or [g], while the long-VOT ("Dave") stimuli were identified as [p], [t], and [k]. Although these eight listeners showed a mean improvement of 12.0 percentage points (ranging from 6.6% to 20.8%) in correct identification, an analysis of their response times by ear showed no effect of test, $F(1,7) = 1.14$, $p = .32$, or of Ear, $F(1,7) = 0.22$, $p = .66$, and no interaction between the two, $F(1,7) = 1.83$, $p = .22$.[7] Thus, even successful learning of these stimuli was not sufficient to induce a change in hemispheric processing –the learning apparently had to involve some aspect of talker identification in order to induce a change in lateralization.

4. Discussion

We trained listeners to identify stimuli differing only in VOT as having been produced by two different talkers. Stimulus differences were small enough that both long- and short-VOT tokens were within the same phonetic (voiceless aspirated) category, so listeners were initially at chance to distinguish the two types of tokens. However, after 3 h of identification training with trial-level feedback, eight listeners showed a clear increase in proportion correct (at least 5 percentage points). These listeners also showed a decrease in response time for correct responses to tokens on which they were trained, but an increase in the time of correct responses to tokens in the generalization set.

4.1. Training

The observed increase in response time for tokens on which listeners were not trained, combined with a decrease in RT for trained tokens, is consistent with an active model of perceptual cue processing in which both bottom-up (data-driven) and top-down (hypothesis-driven) processes interact (e.g., Goldinger & Azuma, 2003; Grossberg, 2003; Nusbaum & Magnuson, 1997). According to such models, speech perception involves the formation of a correspondence between low-level sensory information and high-level knowledge of phonetic categories. This correspondence, in turn, depends on the discovery of reliable sensory features because processing resources are limited and devoting attention to irrelevant features greatly increases the time needed for making accurate decisions (Ahissar & Hochstein, 2004). In the present task each stimulus contains many acoustic properties, including both the categorization-relevant feature of VOT and also task-irrelevant features such as formant transitions and steady-state frequencies associated with consonant place of articulation and vowel identity. Prior to training, listeners ignored VOT difference between long- and short-VOT tokens because both were drawn from within a single phonetic (voiceless, unaspirated) voicing category. When training requires listeners to discover a new distinction they must search through the acoustic signal to discover identifiable differences between stimuli. Thus, in the initial stages of training, listeners attend to more acoustic features than usual on each trial in order to discover those that are reliable cues to talker identity. Attending to more features demands the commitment of more processing resources, causing initially longer response times. Eventually, successful learners discover cues that are consistently useful for making correct responses. By focusing attention on these cues and ignoring other, less reliable acoustic features on any given trial, they are able to reduce demand for limited attentional resources and subsequently reduce response time.

In the current study, the observation that otherwise successful learners do not show a decrease in RT for generalization-set stimuli suggests that they have not (yet) learned to use VOT (alone) as a cue to talker identity. Instead, they may be using some combination of acoustic properties including VOT along with other features, perhaps those typically associated with place of articulation and/or vowel category. We argue that listeners in this experiment are making talker-identity judgments on the basis of a combination of acoustic properties including, but not limited to, VOT. This composite of acoustic features is effective for the trained stimuli, but because it incorporates properties that are specific to them (e.g., features related to their specific

[6] A total of 12 participants were trained in the phonetic identification condition, eight of whom achieved the required improvement of at least 5 percentage points overall (on training set and generalization set tokens combined).

[7] The lack of a significant effect of ear or an interaction of ear with test suggest that training did not lead to a greater right-ear/left-hemisphere advantage for these listeners. The lack of any apparent increase in left-hemisphere processing for the phonetically trained listeners might be explainable in terms of a ceiling effect: Even before training listeners may have been processing these VOT differences as phonetic cues with a left-hemisphere dominance, and therefore increased phonetic training did not lead to any change in lateralization. However, neither group of learners (phonetically trained or talker trained) showed a significant right-ear/left-hemisphere advantage on the pretest, as would be expected under this assumption. One explanation for the lack of an observable left-hemisphere advantage for phonetic processing of VOT is that, all else being equal, non-prototypical exemplars of speech sounds tend to induce weaker neural responses than do more prototypical ones (e.g., Näätänen et al., 1997), and poorer lateralization effects (Blumstein, Meyers, & Risseman, 2005), perhaps because of weaker activation of long-term memory representations of phonological categories specifically in the left hemisphere (Shestakova et al., 2002). All of the stimuli in the present experiment were intentionally poor exemplars of the voiceless unaspirated phonemes of English (to allow for learning), and therefore might have elicited only a weak (insignificant) left-hemisphere advantage. The same (lack of) effect is not observed in the right hemisphere-dominant task of talker identification because, prior to training, listeners do not have any categorical representation for either talker. Training induces the development of (weak) categories, resulting in a (weak) right-hemisphere advantage when these new representations are invoked for talker identification.

place of articulation), when listeners attempt to evaluate this newly learned, composite property in the generalization stimuli, they are unsuccessful. This unsuccessful search for expected (but absent) features consumes attentional resources and increases response times. In support of this hypothesis, it may be observed that generalization tokens differing from the training-set tokens only in terms of consonant place of articulation (/ka/ and /ta/ tokens) showed a much greater increase in RT from pre-training test to post-training test (326 ms) than did those differing only in vowel (/pi/, /pu/) (84 ms). Thus, it is not merely lack of exposure to generalization stimuli that incurs a processing disadvantage. Stimuli that differ from the training set according to consonantal features are more disadvantaged than those differing only according to vocalic features, suggesting that the composite properties listeners are attending to include consonantal, but not vocalic, features. In recent studies of phonetic learning, the most successful results have been achieved using a high-variability training paradigm in which listeners are exposed to multiple tokens from multiple talkers. Researchers argue that such variety allows listeners to discover and direct attention toward more reliable (less context-specific) cues and away from less reliable ones (Bradlow, Pisoni, Yamada, & Tohkura, 1997; Clopper & Pisoni, 2004; Lively, Logan, & Pisoni, 1993; Logan, Lively, & Pisoni, 1991; see Iverson, Hazan, & Bannister, 2005 for discussion), and it seems likely that training with a wider variety of tokens, perhaps over a longer period of time, would have improved the present results as well.

4.2. Lateralization

The present results support a functionally based model of lateralization. Listeners trained to use VOT as a cue to talker identity subsequently increased activation of right hemisphere-localized neural networks, presumably those that have been shown by previous researchers to be involved in processing related to talker identification (Belin & Zatorre, 2003; Belin et al., 2002; Imaizumi et al., 1997; Levy et al., 2001, Levy, Granot, & Bentin, 2003; Nakamura et al., 2001; Van Lancker et al., 1989; Von Kriegstein et al., 2003; Von Kriegstein & Giraud, 2004). Our results are consistent with research suggesting that the relative hemispheric activation for processing particular speech sounds is determined by the task in which the listener is engaged (talker identification vs. phonetic categorization). Gandour and colleagues (Gandour & Dzemidzic et al., 2003; Gandour & Xu et al., 2003) have shown that lateralization of prosodic cue processing is determined at least in part by a top–down, attentionally guided mechanism and that functional specialization emerges from the interaction of functional demands and stimulus properties. Here, we have shown that the same may be true for a *segmental* acoustic cue: training induced listeners to attend to a segmental difference as a cue to talker identity, resulting in a shift in hemispheric lateralization for processing the same sounds.

While we argue that the present results and those of Gandour and colleagues suggest that listeners are changing the *way* they process specific acoustic cues, they are also consistent with the hypothesis that listeners accomplish different tasks (talker identification vs. phoneme recognition) on the basis of *different* acoustic cues, and that it is the change in the type of cue being processed that induces the shift in lateralization. Although we shall argue that aspects of our results do not support this second hypothesis, it cannot be conclusively ruled out at the present time, and therefore it is instructive to consider the implications of the present results for models of lateralization based on stimulus properties rather than task demands.

With respect to stimulus-based models of lateralization, the present results would lend more support to a model in which lateralization is determined by the size of the temporal windows of analysis (e.g., Boemio et al., 2003; Poeppel, 2003) rather than one based on a difference between the analysis of temporal and spectral properties (e.g., Zatorre & Belin, 2001; Zatorre et al., 2002). Since the long-VOT and short-VOT stimuli were spectrally identical, the shift in hemispheric dominance cannot be explained in terms of a shift away from temporal-cue processing and/or toward spectral-cue processing.

In terms of an asymmetric temporal sampling model (Boemio et al., 2003; Poeppel, 2003), the present results could be interpreted as reflecting a shift in listeners' use of shorter (left hemisphere-dominant) cues (e.g., VOT) in favor of cues with longer (right hemisphere-dominant) overall durations (e.g., ratio of VOT to vowel length). Since all properties of the syllables other than VOT and vowel length (adjusted to maintain the total duration of each syllable) were held constant, such a ratio could, in principle, serve as a diagnostic cue for the present task just as well as VOT. Since syllable durations were roughly constant across vowel and consonant classes, it is possible that listeners were able to incorporate vowel or syllable duration as part of a cue to talker identity without incorporating different (spectral) vowel qualities. That is, it would have been possible to use the duration of an [a] just as effectively as that of an [i] without incorporating any properties of the vowel specific to [i] as opposed to [a] in their decision.

Thus, it is possible that training encouraged listeners to shift from using one cue (VOT) to another involving vowel duration either alone or in conjunction with VOT. According to this hypothesis, listeners might initially have focused on VOT as an obvious cue for distinguishing these stimuli but during training they discovered that responding according to their initial (English-based) category boundaries resulted in a large number of incorrect responses and therefore learned to distrust their reliance on VOT as a cue (cf. Francis et al., 2000; for another case in which listeners learned to avoid making responses based on a familiar cue without necessarily resulting in increased use of a weaker cue). Having learned to avoid relying on VOT listeners would need to identify some other cue to use; successful learners were those who identified some other diagnostic

cue, for example consonant/vowel or consonant/syllable duration ratio, or even just overall vowel length, and learned to use this as their cue to talker identity. This interpretation is partially supported by the pattern of results obtained from those subjects in the phonetically trained group who showed successful learning but no shift in hemispheric dominance, suggesting that they maintained their reliance on a short-term (left hemisphere) feature such as VOT. However, this explanation is partly inconsistent with the results of the analysis of responses to the generalization set stimuli, showing that learning was limited to syllables beginning with particular consonants ([t] and [k]). Listeners *must* have continued to use some consonantal properties in their categorization decisions, although it is possible that they *also* incorporated some aspect of vowel duration in their decision following training.

More importantly, accepting an interpretation of the present results in terms of an asymmetric temporal sampling model still begs the question of *why* listeners in the talker-trained group may have chosen to shift their attention to a longer-duration (right hemisphere) cue while those in the phonetically trained group maintained their reliance on a shorter (left hemisphere) property of the stimuli. The only difference between the two groups was the way in which they were instructed to identify the two categories of sound, either as productions of two different talkers or as different phonemes. Thus, the present results suggest that the three-way relationship between the left hemisphere, short-term acoustic events, and phonetic segment identification, and that between the right hemisphere, longer-term acoustic events, and talker identification, must be deeper than mere coincidence. Listeners who were trained to identify talkers changed their hemispheric dominance for processing certain speech sounds in a fundamentally different way than did those who were trained to identify phonetic segments.[8]

On the basis of the present evidence we cannot determine whether this difference involved directly shifting the preference for processing VOT from the left (phonetic) hemisphere to the right (talker) hemisphere, e.g., increasing the involvement of talker identification-specific regions (cf. Imaizumi et al., 1997), or whether the switch was mediated by a change in the acoustic cue(s) used for processing the sounds, shifting from processing VOT in the left hemisphere to processing some cue of longer duration in the right hemisphere. However, in either case the observed shift toward a left-ear/right-hemisphere advantage for processing talker identity is most consistent with a model of speech perception in which processing of acoustic cues is lateralized according to the task (talker identification vs. phonetic categorization) rather than (or in addition to) the acoustic properties used in classification.

4.3. Future research

It is important to note that the prosodic cues used by Gandour and Dzemidzic et al. (2003) were probably approximately equally good at cuing the respective linguistic and non-linguistic qualities. Fundamental frequency is known to be a highly salient cue for both emotional and tonal classification. In contrast, the VOT cue used here is almost certainly a much stronger phonetic cue than it is a cue to talker identity. Although Allen, Miller, and DeSteno (2003) showed that VOT varies reliably according to talker, phonetic category-related differences in VOT are larger (about 50–70 ms between long- and short-VOT categories, as opposed to 1–50 ms between different talkers' productions of long-VOT tokens). Furthermore, English long- and short-VOT categories lie on opposite sides of a discontinuity in auditory sensitivity, making the perceptual distance between them even larger, and affecting learning (e.g., Holt et al., 2004). Conversely, spectral properties (including f0) are typically considered to be the strongest cues to talker identity (e.g., Lavner, Gath, & Rosenhouse, 2000). Thus, the present study may represent a particularly stringent test of changing lateralization, because VOT may not be a very good candidate for talker identification even under ideal circumstances (cf. Allen & Miller, 2004). It may be the case that training-induced changes in laterality are easier to observe using cues that are more equipotent, for example, vowel formant frequencies that have been shown to serve both as primary cues to vowel identity (Shepard, 1972) and also as strong cues to talker identity (Lavner et al., 2000). Future research in this area should focus on stimuli differing in terms of more balanced cues.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*, 457–464.

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*, 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*, 544–552.

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

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Brain Research: Cognitive Brain Research, 13*, 17–26.

Bever, T. G. (1971). The nature of cerebral dominance in speech behavior of the child and adult. In R. Huxley & E. Ingram (Eds.), *Language Acquisition: Models and Methods*. New York: Academic Press.

Blumstein, S. E., Meyers, E. B., & Risseman, J. (2005). The perception of voice onset time: an fMRI investigation of phonetic category structure. *Journal of Cognitive Neuroscience, 17*(9), 1353–1366.

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2003). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience, 8*, 389–395.

Boersma, P., Weenink, D. (2005). Praat (Version 4.2) [Computer software]. <www.praat.org./> [Confirmed March 10, 2006.].

[8] Interestingly, the observation that listeners showed no strong right-ear/left-hemisphere advantage either on the pretest or following phonetic categorization training is consistent with Blumstein et al. (2005) ('s) hypothesis that lateralization may also be related to stimulus prototypicality since the tokens used in the present experiment were all highly non-prototypical, at least with respect to voicing.

Bradlow, A. R., Pisoni, D. B., Yamada, R. A., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Bradshaw, J. L., & Nettleton, N. C. (1988). Monaural asymmetries. In K. Hugdahl (Ed.), *Handbook of Dichotic Listening: Theory, Methods and Research* (pp. 45–69). Chichester, UK: John Wiley and Sons.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47*, 207–239.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics, 67*(2), 224–238.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception and Psychophysics, 62*, 1668–1680.

Friedrich, D. (1974). Comparison of intrusion errors and serial position curves on monaural and dichotic listening tasks: a developmental analysis. *Memory and Cognition, 2*(4), 721–726.

Gandour, J., Dzemidzic, M., Wong, D., Lowe, M., Tong, Y., Hsieh, L., et al. (2003). Temporal integration of speech prosody is shaped by language experience: an fMRI study. *Brain and Language, 84*, 318–336.

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., et al. (2004). Hemispheric roles in the perception of speech prosody. *Neuroimage, 23*, 344–357.

Gandour, J., Wong, D., Lowe, M., Dzemidzic, M., Satthamnuwong, N., Tong, Y., et al. (2002). A cross-linguistic fMRI study of spectral and temporal cues underlying phonological processing. *Journal of Cognitive Neuroscience, 14*, 1076–1087.

Gandour, J., Xu, Y., Wong, D., Dzemidzic, M., Lowe, M., Li, X., et al. (2003). Neural correlates of segmental and tonal information in speech perception. *Human Brain Mapping, 20*, 185–200.

Giraud, A. L., & Price, C. J. (2001). The constraints functional neuroimaging places on classical models of auditory word processing. *Journal of Cognitive Neuroscience, 13*, 754–765.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics, 31*, 305–320.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics, 31*, 423–445.

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: implications for speech perception. *Journal of the Acoustical Society of America, 116*(3), 1763–1773.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., et al. (1997). Vocal identification of speaker and emotion activates different brain regions. *NeuroReport, 8*, 2809–2812.

Indefrey, P., & Cutler, A. (2005). Prelexical and lexical processing in listening. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (3rd ed., pp. 759–774). Cambridge, MA: MIT Press.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: a comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America, 118*, 3267–3278.

Kirk, R. E. (1995). *Experimental design* (3rd ed.). Pacific Grove: Brooks/Cole.

Kreiman, J., Van Lancker-Sidtis, D., & Gerratt, B. R. (2005). Perception of voice quality. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 338–362). Malden, MA: Blackwell.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication, 30*, 9–26.

Levy, D. A., Granot, R., & Bentin, S. (2001). Processing specificity for human voice stimuli: electrophysiological evidence. *NeuroReport, 12*, 2653–2657.

Levy, D. A., Granot, R., & Bentin, S. (2003). Neural sensitivity to human voices: ERP evidence of task and attentional influences. *Psychophysiology, 40*, 291–305.

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

Lively, S. E., Logan, J. D., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*, 1242–1255.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: a first report. *Journal of the Acoustical Society of America, 89*, 874–886.

Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 351–377). Baltimore: York Press.

Nakamura, K., Kawashima, R., Suigiura, M., Kato, T., Nakamura, A., Hatano, K., et al. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia, 39*, 1047–1054.

Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., et al. (1997). Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature, 385*, 432–434.

Nusbaum, H., & Magnuson, J. (1997). Talker normalization: phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Pscychology: Human Perception and Performance, 8*(2), 297–314.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication, 41*, 245–255.

Raphael, L. J. (2005). Acoustic cues to the perception of segmental phonemes. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 182–206). Malden, MA: Blackwell.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532.

Rimol, L. M., Eichele, T., & Hugdahl, K. (2006). The effect of voice-onset-time on dichotic listening with consonant–vowel syllables. *Neuropsychologia, 44*, 191–196.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002a). *E-Prime (Version 1.1) [Computer software]*. Pittsburgh, PA: Psychology Software Tools Inc..

Schneider, W., Eschman, A., & Zuccolotto, A. (2002b). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002c). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Schwartz, J., & Tallal, P. (1980). Rate of acoustic change may underlie hemispheric specialization for speech perception. *Science, 207*, 1380–1381.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*, 2400–2406.

Segalowitz, S. J., & Cohen, H. (1989). Right hemisphere EEG sensitivity to speech. *Brain and Language, 37*, 220–231.

Shepard, R. (1972). Psychological representation of speech sounds. In E. E. David & P. Denes (Eds.), *Human communication: a unified view*. New York: McGraw-Hill (pp.).

Shestakova, A., Brattico, E., Huotilainen, M., Galunov, V., Soloviev, A., Sams, M., et al. (2002). Abstract phoneme representations in the left temporal cortex: magnetic mismatch negativity study. *NeuroReport, 13*(14), 1813–1816.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*, 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. *Cognitive Brain Research, 17*, 48–55.

Von Kriegstein, K. V., & Giraud, A.-L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *NeuroImage, 22*, 948–955.

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

Zatorre, R., Belin, P., & Penhune, V. B. (2002). Structure and function of the auditory cortex: music and speech. *Trends in Cognitive Sciences, 6*, 37–46.