ELSEVIER

Neuropsychologia 45 (2007) 1899–1910

NEUROPSYCHOLOGIA

www.elsevier.com/locate/neuropsychologia

Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex

Tyler K. Perrachione [a], Patrick C.M. Wong [b,*]

[a] *Department of Linguistics & Program in Cognitive Science, Northwestern University, Evanston, IL 60208, United States*
[b] *The Roxelyn & Richard Pepper Department of Communication Sciences & Disorders & Northwestern University Institute for Neuroscience, Northwestern University, 2240 Campus Dr. Evanston, IL 60208, United States*

Received 8 August 2006; received in revised form 27 November 2006; accepted 28 November 2006
Available online 26 January 2007

Abstract

Brain imaging studies of voice perception often contrast activation from vocal and verbal tasks to identify regions uniquely involved in processing voice. However, such a strategy precludes detection of the functional relationship between speech and voice perception. In a pair of experiments involving identifying voices from native and foreign language speech we show that, even after repeated exposure to the same foreign language speakers, accurate talker identification is in a large part dependent on linguistic proficiency. These results suggest that a strong integration between the brain regions implicated in voice perception and speech perception accounts for the accurate identification of talkers.

Keywords: Voice perception; Auditory cortex; Language proficiency; Talker identification; Indexical information

1. Introduction

The ability of the human cerebral cortex to extract meaning from the rapid changes in air pressure during speech is a truly remarkable one, seemingly setting us apart from all other species. What is equally remarkable is that during speech perception we not only extract the underlying message, but additionally gain access to an extensive amount of information about the individual who is speaking. Traditionally, a distinction has been drawn between the "linguistic" features of an utterance, which map phonetic structure onto higher-level words and phrases, and its "paralinguistic" or indexical features (Abercrombie, 1967). These paralinguistic features of the speech stream include speakers' gender, relative age, emotional state, and even where they grew up or whether they have a cold. Taken together, the variability of these features from person to person allows listeners to perceive the identity of individual speakers. Such a process of voice perception is generally examined without considering its relation to the process of speech perception. We present evidence from talker identification experiments, as well as a survey of other work examining the interaction between linguistic and paralinguistic information, showing that accurate talker identification abilities are facilitated by linguistic knowledge. Such results suggest that a strong integration between the brain regions implicated in voice perception and speech right and left superior temporal regions – accounts for the accurate identification of talkers by human listeners.

1.1. Neural studies of voice perception

Despite the failure of early dichotic listening studies of the neural bases of voice perception to show lateralization of voice recognition abilities (Kreiman & Van Lancker, 1988; Tarter, 1984), more recent functional brain imaging studies have identified the superior temporal sulci (STS), especially in the right hemisphere, as the major cerebral locus of human voice perception (e.g. Belin, Zatorre, & Ahad, 2002; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). The importance of the right superior temporal region in general to voice perception is now well established, including its role in the perception of vocal affect (see Wong (2002) for a review). Clinical studies of patients with voice perception impairments often reveal right hemisphere damage (Neuner & Schweinberger, 2000; Van Lancker & Canter, 1982). However, the functional significance

* Corresponding author. Tel.: +1 847 491 2416; fax: +1 847 491 2429.
E-mail address: pwong@northwestern.edu (P.C.M. Wong).

0028-3932/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2006.11.015

of simultaneously processing the phonology and meaning of speech during neural studies of voice remains poorly understood, both because it is understandably difficult to disentangle the two when human voices are the unique carriers of speech, and because behavioral evidence specifically associating speech processing with voice perception has been scant.

Functional neuroimaging studies of voice perception have primarily utilized either a stimulus-based method or a task-based method to localize cortical responses to vocal stimuli. In the stimulus-based method, subjects listen to either vocal stimuli (including speech and non-speech vocalizations) or non-vocal stimuli, such as environmental or animal sounds. Belin, Zatorre, Lafaille, Ahad, and Pike (2000) were among the first to demonstrate increased activation in the superior temporal sulci, especially in the right hemisphere, for vocal versus non-vocal sounds. Fecteau, Armony, Joanette, and Belin (2004) showed that these regions responded preferentially to human vocalizations (speech and non-speech) versus either the vocalizations of animals or non-biological sounds. However, because the tasks in these studies were to listen passively to the stimuli – and because the vocal stimuli included multilingual speech – the results from these stimulus-based methods of identifying the cortical bases of voice perception cannot tell us about how these putative voice-selective regions are functionally engaged in higher cognitive tasks such as talker identification, much less about the functional relationship between speech and voice processing.

Other studies have investigated voice perception using a task-based method, in which the stimuli are the same but subjects make judgments about either the voice or another aspect of the stimulus, such as its verbal content. This method allows the identification of areas that are functionally distinct for processing voices versus other content. For example, Stevens (2004) revealed increased activation in several areas, including left anterior superior temporal gyrus and right medial frontal gyrus, when subjects were attending voices versus words in a working memory "two-back" task. Two perceptual studies of voice recognition (von Kriegstein & Giraud, 2004; von Kriegstein et al., 2003) implicated the right superior temporal sulcus in recognizing target voices versus sentences. There remains a major concern when interpreting the results of such task-based voice perception studies; although they can illustrate regions that respond preferentially to voices versus verbal content, they cannot show the activation common to both tasks, thus obfuscating any functional relationship between them. For example, although Stevens (2004) showed increased activation in the anterior STG for the voice relative to verbal task, the verbal task alone still activated this region significantly more than at baseline. To that extent, even though task-based brain imaging studies of voice perception have revealed areas uniquely implicated in processing voices, they cannot show to what extent these putative voice-processing areas rely on the cotemporaneous processing of linguistic (phonological) information.

Studies employing alternative methods to ascertain the neural systems specifically involved in voice perception are relatively few. Two PET studies of voice recognition (Imaizumi et al., 1997; Nakamura et al., 2001) implicated a bilateral network including specifically the temporal poles. Likewise, in a related study of speech perception, Wong, Nusbaum, and Small (2004) showed increased bilateral activation in the superior temporal regions when subjects listened to multiple versus single voices, suggesting an expansive auditory network for processing variation among talkers. Belin and Zatorre (2003) also employed an adaptation paradigm to reveal greater right anterior STS activation when listening to multiple versus single voices.

1.2. Behavioral studies of speech and voice perception

Although early theories of speech perception suggested that all non-linguistic information is filtered to other processes, leaving only the fundamental phonological units for linguistic analysis (Licklider, 1952), more recent studies on language have come to appreciate the relevance of variability due to voice. Mullenix and Pisoni (1990) showed that variability in non-attended features of a speech signal interfered with the speed with which individuals responded to the attended features. The authors concluded that the increase in reaction time as stimulus variability increased was a function of talker normalization—the process of adjusting one's phonetic perception to the idiosyncrasies of an individual's voice. Likewise, Johnson (1990) showed that when acoustic features of a vowel were held constant, but subjects believed the vowel was produced by different talkers, they judged the quality of the vowel to be different. More recently, Allen and Miller (2004) showed that features as brief as individual differences in voice-onset time for stop consonants could affect talker-specific processing in speech perception. Other studies have shown that inter-talker variability affects not only the perception of speech, but also how it is encoded in memory. Bradlow, Nygaard, and Pisoni (1999) and Palmeri, Goldinger, and Pisoni (1993) showed that individuals were better at recognizing that they had heard a word previously when it was spoken by the same talker rather than different talkers. Nygaard and Pisoni (1998) took a related approach to showing that the phonetic characteristics of individual talkers were retained in memory for use in speech perception. In their study, subjects who had practiced listening to a speaker found novel sentences spoken by the same speaker more intelligible when embedded in noise than did subjects who had no prior experience with the speaker.

These studies provide strong evidence that the so-called "paralinguistic" information is not "filtered" out during speech perception, and that listeners are sensitive to the acoustic nuances of individual talkers even during a "linguistic" process (perceiving the phonemic, and ultimately lexical-semantic content of the acoustic signal). These acoustic nuances could be encoded as episodic memory traces for future speech perception (Goldinger, 1997; Pisoni, Saldana, & Sheffert, 1996), or they could represent forms of lawful variability that listeners actively and dynamically adapt to during speech perception (Nusbaum & Magnuson, 1997). It is possible that listeners' awareness of the phonetic differences between talkers for the purpose of speech perception could likewise be used for the process of talker identification. However, note that in order for listeners' to be sensitive to the phonetic differences between talkers, they must also have perceptual access to the

relevant distinctions. Individuals' perception and encoding of the phonemic information necessary for talker normalization and perhaps talker identification is strongly influenced by their language experience. There is a substantial literature showing that language familiarity fundamentally affects whether individuals perceive differences between two sounds as variation within or across phonetic categories (e.g. Best, 1994; Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Liberman, Harris, Hoffman, & Griffith, 1957; Werker & Tees, 1984), to the extent that different neural mechanisms subserve the perception of native and non-native speech sounds (e.g. Wong, Parsons, Martinez, & Diehl, 2004). Additionally, Samuel (1997) and Ganong (1980) showed that consonants along a voicing continuum were more likely to be categorized in such a way that they formed real words, suggesting top–down influence in identifying marginal exemplars. Only individuals who have some degree of proficiency in a language will be able to perceive and encode the sounds of that language in a meaningful way, and thereby gain access to the inter-talker phonetic variability necessary for accurate speech and voice perception.

Two early studies (Goggin, Thompson, Strube, & Simental, 1991; Thompson, 1987) looked at the influence of language on voice identification. In these studies, subjects were presented "voice line-ups" in which they were to pick a target voice form among a sequence of distracter voices. After brief familiarization to a target voice, Thompson (1987) showed that native English listeners more often correctly selected the target from the line-up when listening to English compared to Spanish. However, no comparison condition existed to assess the accuracy of native Spanish speakers in identifying the voices, and the underlying cause of the effect was not pursued. Using the same methodology, Goggin et al. (1991) extended those results to show that English listeners were relatively impaired in recognizing voices speaking German, while German listeners were likewise impaired in recognizing voices speaking English. Although the experiment was not designed to test a specific theoretical framework, Goggin and colleagues' post hoc interpretation of the results proposed a schematic representation of voice, wherein increased ability to identify one set of voices is based on increased exposure to those voices relative to others. However, the authors' results do not reveal whether such schemata differentially represent exposure and specific linguistic knowledge. Without uniquely identifying the role of linguistic knowledge, schema-based accounts necessarily predict that sufficient exposure to a novel set of voices will be sufficient to allow accurate voice recognition. In the present study we aim to demonstrate that exposure alone is insufficient for native-like voice identification, suggesting that linguistic knowledge specifically underlies the language familiarity effect.

1.3. Contribution of the present study

Despite consistent implication in brain imaging studies of voice, the precise role of the left superior temporal cortex and other language areas in voice perception remains poorly understood. Because of the methods used, authors of voice perception studies have been unable to demonstrate a functional interaction between regions previously implicated in language and speech perception and those they found for voice perception (e.g. von Kriegstein et al., 2003). Although such studies have revealed regions of increased activated for voice-processing tasks, it is as yet unclear how these interact with other areas simultaneously activated while processing the latent phonological and semantic information in the signal. The biggest impediment we see to saying with certainty how the left superior temporal sulcus and other language areas are also integrated into a voice-processing network is the fact that there has been little behavioral evidence linking speech perception to voice perception ability. We present such evidence in the two experiments below.

We adopt the position that without language familiarity the relevant phonetic information contributing to accurate talker identification is unavailable, and thus listeners will be unable to identify voices speaking in a foreign language as accurately as in a familiar one. However, given a sufficient amount of familiarity with the phonology of a non-native language, individuals should be able to learn to identify voices in that language as accurately as in their native one. In the following pair of experiments we set out to demonstrate that: (1) Individuals are less accurate at identifying talkers speaking an unfamiliar language than a familiar one; (2) This effect is truly due to language proficiency. A true linguistic effect on voice perception can be contrasted with more general auditory expertise in talker identification. Such an effect of linguistic proficiency entails that knowledge about a language's phonology contributes to accurately identifying voices speaking in that language. Listeners unfamiliar with a language will not know which acoustic variations signal a meaningful linguistic contrast (e.g. the difference between a voiced and a voiceless consonant), and which variations are attributable to the phonetic idiosyncrasies of an individual talker (e.g. slightly longer voice-onset times within the same phonetic category, Allen & Miller, 2004). Because being able to discern individual characteristics from meaningful variation requires phonological knowledge, a true linguistic effect suggests that even repeated exposure will not mitigate any performance gap between voices speaking a familiar versus foreign language. However, if some degree of linguistic information is available, such as in the case of second-language learners, individuals should be able to take advantage of this information to learn more accurate speaker identification.

2. Experiment 1: Single-session speaker identification

Experiment 1 investigated whether familiarity with English and Mandarin affects individuals' abilities to recognize the voices of individuals speaking in these languages.

2.1. Method

2.1.1. Participants

Two groups of listeners participated in Experiment 1. The first group consisted of sixteen native speakers of American English (eleven females, five males, aged 18–26 years, mean = 20.69). None of the English subjects had any familiarity with Mandarin Chinese. The second group consisted of twelve native speakers of Mandarin (eight females, four males, aged 24–31 years, mean = 26.67). All of the Mandarin subjects were born in China and reported speaking exclusively Mandarin Chinese with their parents, though at the time of the experiment, they

were residing in the United States, working as graduate students or research assistants at Northwestern University, or family and friends thereof. Seven of the Mandarin subjects reported that they now used predominately English, though none reported using English for more than 5 years (mean = 2.85). The remaining five Mandarin subjects reported using predominately Mandarin, or using Mandarin and English equally. All subjects reported having no history of hearing or neurological impairments. Each subject gave informed, written consent overseen by the Institutional Review Board at Northwestern University and received a nominal cash payment for participating. Two additional subjects participated in the experiment but were excluded from the analyses because they were familiar with the individuals whose voices were used in the recordings.

2.1.2. Stimuli

Two sets of sentences were recorded for this experiment: one spoken in English, the other in Mandarin (Open Speech Repository, 2005). The English sentences were read by five male native speakers of American English (aged 19–26 years, mean = 21.6). The Mandarin sentences were read by five male native speakers of Mandarin Chinese (aged 21–26 years, mean = 22.6). None of the talkers for the English sentences was the same as for the Mandarin sentences; likewise, none of the individuals recorded as talkers participated in the talker identification experiment. Each talker was assigned a pseudonym used by subjects to identify his voice. Pseudonyms were easily pronounceable, language-appropriate, familiar monosyllabic names, each beginning with a unique letter. The English talker pseudonyms were Bill, Dan, Josh, Mark, and Steve; and the Mandarin talker pseudonyms were Chen, Hong, Liu, Peng, and Wei. Talkers were asked to read the sentences naturally, as though they were having a conversation with a friend. Recordings were made in a sound-attenuated chamber via a SHURE SM58 microphone using a Creative USB Sound Blaster Audigy 2 NX sound card onto a Pentium IV PC. Recordings were sampled at 22.05 kHz using an in-house software Wavax for Windows v2.3 and normalized for RMS amplitude to 70 dB SPL. In all, ten sentences were recorded in each language (see Appendix A). Five of these were selected as "training sentences" and the remaining five as "generalization sentences".

2.1.3. Procedure

Experiment 1 had two language conditions, in which sentences were spoken by either English talkers or Mandarin talkers. The order of conditions was counterbalanced across subjects to mitigate effects of task familiarity. Each condition consisted of a familiarization phase and a final talker identification test, and subjects participated in both language conditions during a single experimental session, which lasted about 40 min. Subjects were tested individually in a sound-attenuated chamber. Before the experiment began, subjects were instructed that they would be learning to recognize five male talkers by the sound of their voice. For both the English and Mandarin language conditions, the subjects were told that what the talkers were talking about was not important, and it did not matter whether they could understand what was being said; the only thing that they needed to pay attention to was *who* was talking.

During the familiarization phase, subjects practiced identifying the five talkers of one language. One talker's name would appear on the screen while a recording of him saying a sentence was played over the headphones. After the listener had heard the first talker, the next talker's name would appear while a recording of him reading the same sentence was played. After the listeners heard all five talkers in this way, they took a short quiz with feedback over the sentence they had just practiced. During the quiz, all five talkers' names would appear on the screen at the same time, while one of them read the sentence. Subjects identified which talker they believed was speaking by pressing a corresponding button on a computer keyboard. If they answered correctly, the computer told them they were correct. If they answered incorrectly, the computer told them who the correct speaker was. After the subject finished practicing recognizing speakers on one sentence, they performed the same familiarization—practice quiz task on each of the remaining four training sentences. The order of training sentences, as well as the order of talkers within blocks, was randomized across subjects.

After the subject had been familiarized with all five talkers saying each of the five training sentences, they underwent a final talker identification test. This test consisted of all 25 training tokens, and the procedure was the same as the practice quizzes except no feedback was given. Additionally, the 25 generalization tokens (five talkers × five "generalization sentences") were included in the final talker identification test. The 50 test tokens were presented randomly for each subject. Only accuracy on this final talker identification test, and not practice quizzes, was used to gauge subject performance for Experiment 1. After completing one language condition, the subjects repeated the entire procedure for the other language condition. This procedure is similar to our previous non-native speech learning study (Wong & Perrachione, in press; Wong, Perrachione, & Parrish, in press).

2.2. Results

Subjects' accuracy on Experiment 1 was measured as the number of tokens they correctly identified during the final talker identification test out of the total tokens presented. We examined whether the native language of the listeners, the language being spoken in the stimuli, or whether the stimuli were those used in training had any effect on subjects' accuracy scores. Subjects' performance on the various conditions is illustrated in Fig. 1. A three-way repeated-measures ANOVA with Condition (English talkers versus Mandarin talkers) and Stimulus Practice (training sentences versus generalization sentences) as within-subject factors and Subject Group (English subjects versus Mandarin subjects) as a between-subject factor was conducted on the accuracy measures from Experiment 1. This test revealed no main effect of Condition [$F(1, 26) = 1.845$, $p = .187$], indicating that neither set of talkers was overall easier to identify, nor of Practice [$F(1, 26) = .216$, $p = .646$]. There was a significant Condition by Subject Group interaction [$F(1, 26) = 25.259$, $p < .001$], indicating that Mandarin subjects were more accurate at Mandarin talkers, and English subjects at English talkers (see Fig. 1). There was also a significant three-way interaction between Subject Group, Condition, and Practice [$F(1, 26) = 18.477$, $p < .001$], which suggests that subjects' ability to generalize in talker identification to the unpracticed sentences was only affected when they were familiar with the language. The interaction effect between Practice and Subject Group was not significant [$F(1, 26) = 1.038$, $p = .318$], suggesting that both groups were equally likely to generalize in voice identification to unpracticed sentences. Likewise, the interaction effect between Condition and Practice was not significant [$F(1, 26) = .884$, $p = .356$], suggesting that subjects were able to generalize talker identification on both the English and Mandarin sentences.

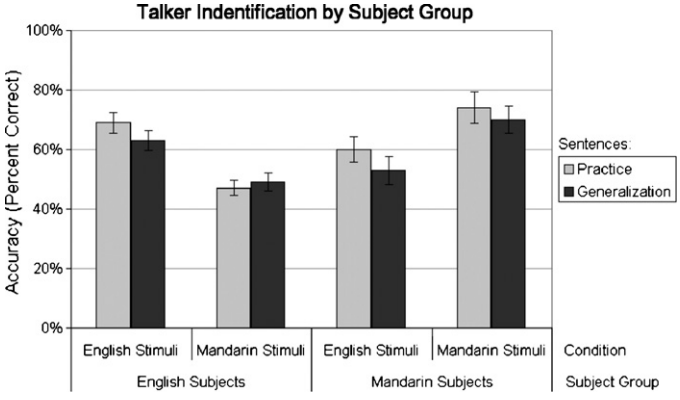

Fig. 1. Mean talker identification performance by English and Mandarin subjects in each language condition of Experiment 1. Error bars indicate the standard error of the mean.

We found the lack of a significant main effect of Practice to be surprising, since individuals generally test better on materials they have practiced than those they have not. This was especially curious because of the significant Subject Group by Condition by Practice interaction. Post hoc analyses comparing subjects' accuracy on the practiced "training sentences" versus the novel "generalization sentences" were conducted for the English subjects in the Mandarin language condition and Mandarin subjects in the English language condition. Paired-sample *t*-tests revealed that English subjects performed equally well on identifying the practiced and novel Mandarin sentences [$t(15) = -.823$, $p = .424$], but that Mandarin listeners performed significantly better on the practiced items [$t(11) = -2.765$, $p < .05$, corrected]. This suggests that the significant Subject Group by Condition by Practice interaction seen in the original ANOVA was driven by the lack of a practice effect for English subjects listening to Mandarin talkers.

2.3. Discussion

The results from Experiment 1 reveal a clear effect of language familiarity on individuals' ability to identify talkers. Foremost, individuals are better at identifying talkers speaking in their native language than in a second- or non-native language. This experiment additionally shows that because performance on the generalization items was above chance (20 percent) in all conditions, individuals are able to generalize talker identification to novel utterances, even in a non-native language. However, this ability to generalize to unpracticed utterances was most often worse than practiced utterances. Although such practice effects could be attributed to either an increased ability to identify talkers from identical utterances after practice or to some inherent difference in the difficulty of the two sets of sentences, the observed difference in performance still underscores the importance of utterance content when designing voice recognition paradigms, such as forensic "voice line-ups" or clinical tests of voice recognition impairment. Additionally, because the effect of practice was not seen for English subjects attending Mandarin talkers, it remains an open question to what extent individuals with no familiarity in a non-native language can benefit from practicing identifying talkers in that language. The issue of non-native language practice is addressed further in Experiment 2.

The question remains whether the differences in accuracy between subject groups on the English and Mandarin language conditions is in fact due to language proficiency (a true linguistic effect), or whether it is due to a more general lack of experience identifying voices speaking an unfamiliar language (a task effect). Although these two possibilities seem similar, there are distinct and theoretically relevant claims underlying each. People spend a great deal of time listening to voices speaking their native language and associating them with the individual talking, but it is unlikely that people will ever devote a substantial amount of time to doing this in a language with which they are unfamiliar. In Experiment 2 we address the question of whether the underlying cause of the language familiarity effect seen here and in earlier studies (Thompson, 1987; Goggin et al., 1991) can be attributed specifically to linguistic proficiency, or whether factors other than linguistic knowledge are in play. Short-term training programs have often been shown to improve performance on auditory perception tasks to native-like levels within the laboratory context (e.g. Golestani & Zatorre, 2004; Wong & Perrachione, in press). Experiment 2 utilizes such a short-term training study for voice identification in both familiar and less-familiar languages. If the underlying cause of the language familiarity effect is linguistic proficiency (a true linguistic effect), then the magnitude of the effect should remain constant across the course of training for individuals completely unfamiliar with the foreign language. Additionally, individuals with some degree of second-language proficiency may be able to learn to utilize their knowledge about the non-native language for use in voice perception. However, if the underlying cause of the language familiarity effect is not truly linguistic in nature, but has more to do with exposure to voices speaking a foreign language (that is, specific linguistic knowledge is not necessary), then the magnitude of the language familiarity effect should attenuate across the course of the training program, as individuals' exposure to voices speaking the foreign language increases.

3. Experiment 2: Short-term talker identification training

Experiment 2 examined whether short-term training at identifying voices speaking either a native language or a non-native language could overcome the language effect on identification accuracy. If individuals with no familiarity with a non-native language can learn to identify voices in that language after training as accurately as in their native language, then the observed language effect in talker identification is a product of a more general voice perception system; whereas if the accuracy gap remains after substantial training, the language effect is a product of an interaction between speech and voice perception systems for accurate talker identification by human listeners.

3.1. Method

3.1.1. Participants

Similar to Experiment 1, two groups of subjects participated in Experiment 2: an English subject group and a Mandarin subject group. No talker who participated in recording the stimuli or subject from Experiment 1 participated also in Experiment 2. The English subject group consisted of thirteen subjects (four male, nine female, aged 18–26 years, mean = 19.62). None of the English subjects had any familiarity with Mandarin Chinese. The Mandarin subject group consisted of thirteen subjects (three males, ten females, aged 19–39 years, mean = 25.46). The Mandarin subjects were all residing in the United States at the time of the experiment, either as students or employees of Northwestern University or family and friends thereof, and all had functional English language skills. All subjects reported having no history of hearing or neurological impairments. Each subject gave informed, written consent overseen by the Institutional Review Board at Northwestern University and received a nominal cash payment for participating in each training session. If a subject completed all training sessions, he or she received an additional cash bonus. One additional subject completed the experiment but was excluded from the analyses because her responses were sometimes too slow to be accurately logged by the stimulus presentation software.

3.1.2. Stimuli

The stimulus sentences in this experiment are the same as in Experiment 1. During the training portion of this experiment, however, only the five "training

sentences" were used for both familiarization and testing. The five "generalization sentences" were reserved for an additional Generalization Test that was performed at the end of the entire training program.

3.1.3. Procedure

The procedure for Experiment 2 extended the single-session design of Experiment 1 into a 6-day training program. The number of training sessions was set at six because early experiment pilots suggested that subjects' performance reached a plateau around this point. Subjects participated in one 30-min training session per day, and care was taken to make sure the entire sequence of six training sessions was completed within the time frame of 8 days.

Experiment 2 consisted of two language conditions, with either English talkers or Mandarin talkers. Each condition was made up of two phases, a familiarization phase and a testing phase. The format of the familiarization phase was identical to the equivalent phase in Experiment 1. The testing phase of Experiment 2 consisted of only the "training sentences" that subjects practiced during familiarization, and unlike Experiment 1, it did not contain a generalization component. Otherwise, the format for the testing phase of Experiment 2 was identical to that of Experiment 1. The order of conditions was counter-balanced across subjects, and subjects always participated in both conditions in each training session. The order of training sentences, as well as the order of talkers within blocks, was randomized across sessions. Subjects' performance on each day of the training program was measured by their accuracy at identifying the talkers during the testing phase of each condition. Accuracy was defined as number of correct identifications out of the total number of tokens presented.

After subjects had completed all 6 days of the training program, they performed an additional generalization test. The format of this program-final Generalization Test was identical to that of the final talker identification test from Experiment 1, including both the practiced "training sentences" and novel "generalization sentences." Two subjects (one from each subject group) were excluded from analyses of the Generalization Test because their results were obfuscated by technical problems.

In this experiment, we additionally sought to ascertain whether the two listener groups were finding the same talkers more or less confusable. If the groups were misidentifying the same patterns of talkers, that would suggest they were using similar acoustic cues to talker identity. However, if the groups differed in which talkers they were confusing, it would suggest they were relying primarily on different acoustic cues, or were weighting the relationships between cues differently. For each listener group, we combined their responses into a 5×5 talker identification matrix, such that one diagonal of the matrix corresponded to correct identifications, and the remaining cells represented the number of instances that *talker x* was mistakenly identified as *talker y*. The relationship between these matrices was calculated using the similarity choice model (SCM) (Luce, 1963; Shepard, 1957). Following Clopper and Pisoni (2006), the full similarity choice model was used to calculate the similarity parameters for each matrix, which represent the degree of perceptual similarity between pairs of talkers. Next, the similarity between two comparison matrices was assessed using the restricted similarity choice model. In the restricted model, the similarity parameters between the two matrices are held constant such that for a test comparing those two matrices, the null hypothesis is that the similarity parameters of either individually do not differ significantly from the similarity parameters of both together. However, if the restricted model produced similarity parameters that fit the data worse than the similarity parameters of the full model, then the perceptual similarities mapped by the two matrices differed significantly. Significance was assessed by comparing the G^2_{test} statistic computed from the models' output against a critical χ^2 value, determined by the degrees of freedom of the input matrices.

3.2. Results

3.2.1. Talker identification accuracy

The mean performance of each listener group (English and Mandarin subjects) in each language condition is illustrated in Fig. 2 for each day of training. A three-way repeated-measures ANOVA was conducted on their accuracy scores in the first and last session, with Session (first versus last) and Condition

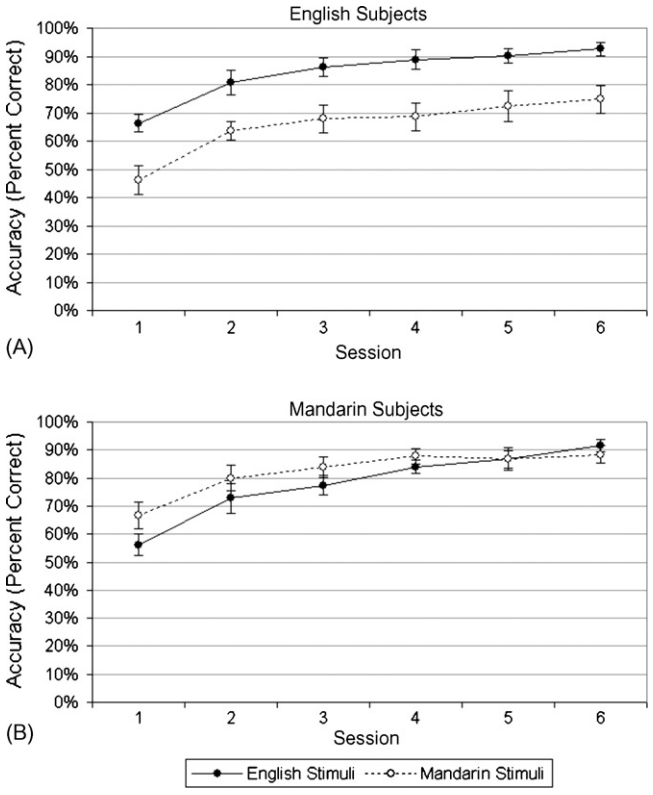

Fig. 2. Mean talker identification performance of the English subjects (panel A) and Mandarin subjects (panel B) in each language condition after each training session of Experiment 2. Error bars represent the standard error of the mean.

(English talkers versus Mandarin Talkers) as within-subject factors, and Subject Group (English subjects versus Mandarin subjects) as a between-subject factor. There was a main effect of Session [$F(1, 24) = 140.166$, $p < .001$], demonstrating improvement across the course of the training program. There was also a main effect of Condition [$F(1, 24) = 9.534$, $p < .01$], indicating better performance in the English condition; however, as will be discussed, this is driven by the fact that the Mandarin subjects performed more like the English subjects on the English talkers over the course of training. The interaction between Subject Group and Condition was also significant [$F(1, 24) = 20.669$, $p < .001$], indicating that subjects were in general more accurate at identifying talkers in their native language than in their non-native language. Additionally, there was a significant three-way interaction between Subject Group, Condition, and Session [$F(1, 24) = 6.006$, $p < .03$], which was driven by the convergence of performance on the Mandarin and English language conditions for the Mandarin subjects across the course of the training program, whereas the performance of the English subjects between the two conditions did not converge. No significant Subject Group by Session [$F(1, 24) = .076$, $p = .785$] or Condition by Session [$F(1, 24) = 3.184$, $p = .087$] interaction effect emerged, suggesting that neither group learned more, nor was either set of sentences in general easier to learn.

3.2.2. Confusability of talkers

We performed eight comparisons between pairs of matrices to determine whether there were similarities in the identification

Table 1
Similarity choice model comparisons for talker identification matrices (Experiment 2)

Comparison #	Talkers	Session	G^2_{test}	Significance (p)
Between listener group comparisons (English subjects vs. Mandarin subjects)				
(1)	English	#1	26.287	*ns*
(2)	English	#6	20.874	*ns*
(3)	Mandarin	#1	96.117	<.001
(4)	Mandarin	#6	78.88	<.001
Comparison #	Talkers	Listener group	G^2_{test}	Significance (p)
Within listener group comparisons (Session #1 vs. Session #6)				
(5)	English	English	191.667	<.001
(6)	English	Mandarin	151.987	<.001
(7)	Mandarin	English	242.875	<.001
(8)	Mandarin	Mandarin	121.477	<.001

For all reported tests, $df = 26$, $\chi^2_{crit} = 38.885$.

patterns, the results of which are delineated in Table 1. We examined between-listener-groups differences (comparing English and Mandarin listeners on one talker language within a given session, testing whether the two groups were making similar errors). We also examined within-listener-group differences (comparing each group on one talker language between sessions, testing whether the error patterns were the same at the beginning and end of training). There was not a significant difference between the patterns of identification of English talkers by the English and Mandarin subject groups in either the first or the last session (comparisons (1) and (2)). Conversely, there was a significant difference between the patterns of identification of the Mandarin talkers by the English and Mandarin subject groups for both the first and the sixth session (both $p < .001$, comparisons (3) and (4)). Overall, there was a significant difference on listener groups' performance for each talker language over the course of the training program, further indicating learning in all conditions (comparisons (5) through (8)).

3.2.3. Program-final generalization test

At the end of the six training sessions, subjects were additionally tested on their ability to generalize talker identification to unpracticed sentences. These data are summarized in Fig. 3. A three-way repeated-measures ANOVA was conducted on the Generalization Test accuracy scores, with Condition (Mandarin versus English talkers) and Practice ("training" versus "generalization" sentences) as within-subject factors, and Subject Group (English versus Mandarin subjects) as a between-subject factor. There was a main effect of Practice [$F(1, 22) = 20.525$, $p < .001$], showing that subjects performed better on the practiced sentences than the novel ones. There was likewise a main effect of Condition [$F(1, 22) = 7.472$, $p < .02$], indicating better performance on the English sentences overall, which, as discussed earlier, was driven by the convergence of the Mandarin subjects' performance in the two conditions over the course of the training program. A significant Subject Group by Condition interaction [$F(1, 22) = 16.611$, $p < .001$] again confirmed that each subject group performed better in their native language than in their non-native language. In the final Generalization Test there was also a significant Practice by Subject Group interaction [$F(1, 22) = 4.999$, $p < .04$] indicating that the Mandarin subjects performed better on the "generalization sentences" than the English subjects. No other interaction effects were significant.

Although subjects' performance on the "generalization sentences" during the program-final Generalization Test remained significantly lower than their performance on the "training sentences," it did show substantial improvement over the ability to generalize after only a single practice session, such as in Experiment 1 (see Fig. 3). These results additionally suggest that the increased performance seen on the "practice sentences" is not due only to over-exposure to the training stimuli, but indeed represents an increased ability in identifying these talkers.

3.3. Discussion

The results from Experiment 2 add strong support to the claim that there is an important linguistic component to accurate talker identification by humans. Like Experiment 1, the English subjects (who had no prior experience with Mandarin) were more accurate at identifying individuals speaking English than those speaking Mandarin. The magnitude of this effect held over the course of the entire training program. The Mandarin

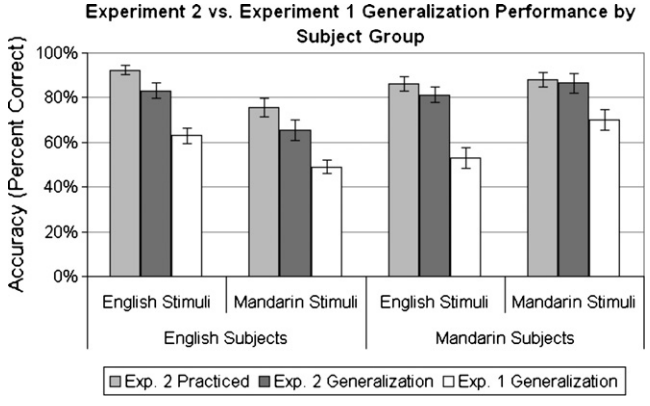

Fig. 3. Mean talker identification performance of the English and Mandarin subject groups in each language condition on the program-final Generalization Test. Performance of the subjects from Experiment 1 on the same "generalization sentences" is included to show the improvement in generalization ability after short-term training. Error bars represent the standard error of the mean.

subjects, meanwhile, performed marginally better in their native language than their non-native language early in the training program, but the effect of language familiarity had all but vanished by the latter sessions. Additionally relevant is that both groups showed similar patterns of errors for the English talkers at both the beginning and end of the training program, but significantly different patterns of errors for the Mandarin talkers. The fact that the Mandarin subjects' patterns of misidentification were never significantly different from those of the English subjects, taken together with their converging performance in the two language conditions, suggests that their familiarity with English allowed them to learn to identify individuals speaking English with native-like adeptness. That is, the Mandarin subjects improved in their ability to use their knowledge of English to encode relevant variation between speakers to identify their voices. Conversely, the English subjects, who had no familiarity with Mandarin, showed different patterns of misidentification from those made by the Mandarin subjects, and their performance in the Mandarin condition improved, but it never approached their performance in the English condition. Additionally, post hoc t-tests conducted on the last three sessions (Session 4 versus Session 5 $t(12) = -2.125$, $p = .055$; Session 5 versus Session 6 $t(12) = -.864$, $p = .404$; Session 4 versus Session 6 $t(12) = -.406$, $p = .406$) revealed no significant improvement of the English subjects' performance at identifying Mandarin voices by the end of the training program. Reaching this plateau suggests that further training would not have improved their performance on this task.

Previous studies of the changes in neural activity over the course of a short-term training program (Golestani & Zatorre, 2004; Wong et al., in press) have revealed that as individuals' ability to perform auditory tasks improved, their patterns of brain activation came to resemble those of experts (i.e., native speakers of a language). In the present study, we showed that although English subjects improved at the task of non-native language speaker identification, they never came to perform as well as native speakers. Because subjects' were not specifically trained on any linguistic features of the test sentences, we propose that the English subjects were unable to gain access to the additional (linguistic) features necessary to close the gap between a familiar and less-familiar language. On the other hand, as suggested by the eventual convergence of their performance in the Mandarin and English conditions, we conclude that the Mandarin subjects were able to use their familiarity with English to learn to use the linguistic cues necessary for optimal talker identification performance.

4. General discussion

Previous brain imaging studies have not been designed to assess the functional role of putative language areas in voice perception despite the observation of activation above baseline in these regions (e.g. Stevens, 2004) in part because the behavioral integration of speech and voice processing had not previously been clearly demonstrated. In this pair of experiments, we have shown a consistent effect of language familiarity on individuals' ability to recognize talkers. Individuals are less able to identify talkers in a non-native language than their native one. After a short-term training program where listeners practiced recognizing the *same five talkers* saying the *same five sentences* for 6 days, only listeners who had some familiarity with the speakers' language were able to reach peak performance. This pattern of results is not simply due to lack of practice with non-native language talker identification, but in fact derives from a true linguistic effect: individuals with no familiarity in the language of a speaker are unable to gain access to the critical linguistic cues facilitating accurate talker identification.

These results relate the neurophysiological results from brain imaging studies of voice perception with behavioral findings from the study of speech perception in a way that suggests these two abilities may be more closely integrated than previously thought. Studies of speech perception have established that individual phonetic variation between talkers can affect how listeners understand or encode an utterance (e.g. Bradlow et al., 1999). Likewise, functional neuroimaging studies of voice perception often reveal activation above baseline in putative language areas of the left superior temporal region, even for voice perception tasks (e.g. Stevens, 2004). The present study provides a first set of behavioral results that may guide future investigation of why these left superior temporal language areas are important during voice perception and, especially, talker identification.

Higher-order cognitive auditory abilities such as talker identification are likely to employ vast networks of brain regions, encompassing extensive networks of lower-level processes from both temporal and spectral domains (Zatorre, Belin, & Penhune, 2002). By better understanding the behavioral processes which human listeners actually employ in identifying voices, we can thereby gain a better understanding of the association of cortical regions that underlie such processes in the brain. This is especially desirable in the context of Small and Nusbaum (2004), who advocate research into neural substrates of behavior that more closely models the most natural setting of that behavior. Similarly, models of person recognition (e.g. Neuner & Schweinberger, 2000) help to understand the behavioral and neural building blocks that contribute to this more abstract higher-level facet of cognition. In an estimable review of the brain bases of voice perception to date, Belin, Fecteau, and Bédard (2004) propose a model of person identification from voice. Their derivation of such a model from models of face perception is likely to help reveal the underlying neural mechanisms of talker identification, since the functional organization of auditory systems is often analogous to that of visual systems (e.g. Alain, Arnott, Hevenor, Graham, & Grady, 2001).

However, the results of the present experiments suggest that such a model of human talker identification be amended to include the bi-directional relation between language and voice. Our present results indicate that accurate voice perception is at least in part dependent on speech perception abilities. As reviewed in the Introduction, previous studies (e.g. Allen & Miller, 2004; Johnson, 1990) have demonstrated that speech perception abilities likewise depend at least in part on information from the voice perception system. Bearing in mind these previous studies in addition to our own, we propose such an integrated

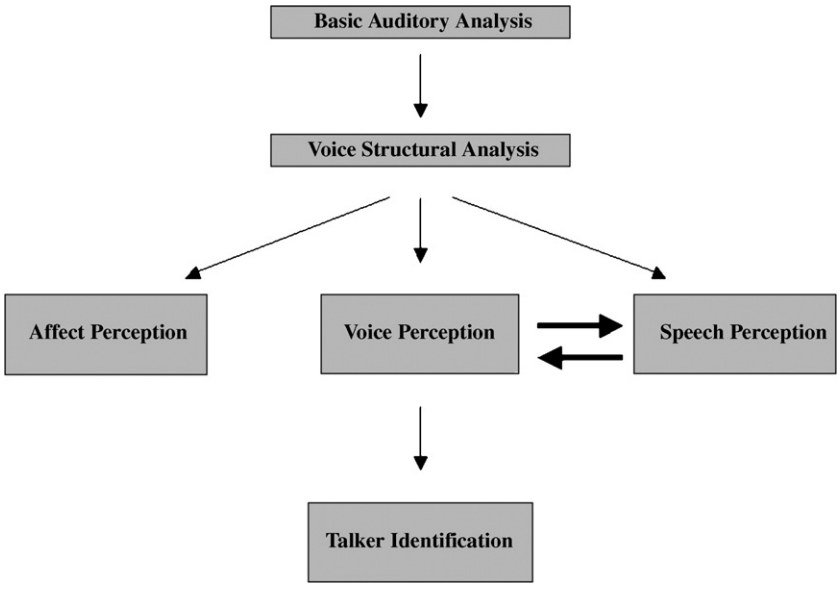

Fig. 4. A model of talker identification by human listeners. The present study suggests the inclusion of bi-directional integration, represented by the bold arrows, associating the processes underlying speech and voice perception. The basis of this model is derived from the voice perception model proposed by Belin et al. (2004).

model, illustrated in Fig. 4, based on the model put forward by Belin et al. (2004).

Our findings that access to the linguistic (phonetic) features of speech contributes to accurate talker identification even after extensive exposure to the same stimuli are consistent with previous studies of talker identification. It has been shown that listeners can learn to recognize talkers of their native language from recordings converted to sinewave speech, which preserves the individual phonetic variation between talkers while removing information about voice quality (Remez, Fellowes, & Rubin, 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002). Likewise, although the specific factors or cognitive processes underlying this effect were not explored, two early studies of non-native language talker identification using "voice line-ups" also showed that listeners were more accurate at identifying voices speaking the same language as themselves (Goggin et al., 1991; Thompson, 1987). The results in the present study have further expanded our understanding of human talker identification by revealing the integrated roles of speech and voice perception.

The language familiarity effect demonstrated here for auditory voice perception appears intuitively analogous to the well-documented own-race bias effect for visual face perception (e.g. Malpass & Kravitz, 1969). In the own-race bias effect, individuals demonstrate a diminished ability to remember and distinguish between faces of another race versus their own race. Much recent work has suggested the cognitive bases for the own-race bias may lie in perceptual learning, in which individuals learn to attend to task-relevant details (see Meissner & Brigman, 2001, for a review). Valentine (1991) proposed a "face-space" model in which the dimensions along which facial features are measured are shaped by prior experience. The own-race bias emerges from this model because the details relevant to identifying faces of different races may be differentially represented across the various perceptual dimensions (e.g. Caldara & Abdi, 2006). Studies investigating interracial contact (e.g. Slone, Brigham, & Meissner, 2000) lend support to this hypothesis, in that individuals reporting greater exposure to the other race also show smaller own-race biases. Likewise, short-term training experiments (e.g. Malpass, Lavigueur, & Weldon, 1973; Elliott, Wills, & Goldstein, 1973) have demonstrated that the magnitude of the own-race bias is attenuated after training on other-race faces. The neural correlates of the own-race bias seem to lie in the fusiform region, which shows differential activation to same- and other-race faces (Golby, Gabrieli, Chiao, & Eberhardt, 2001). Some research suggests this region may also underlie visual expertise in other, non-facial domains (e.g. Gauthier, Skudlarski, Gore, & Anderson, 2000, but cf. Kanwisher, 2000). The relationship between face perception and voice perception is of additional interest given recent findings by von Kriegstein, Kleinschmidt, Sterzer, and Giraud (2005), which suggest face perception regions are engaged during the recognition of familiar voices.

Similarities to the own-race bias in face perception make this a useful analogy to draw when conceptualizing the language familiarity effect. However, there exists an important distinction between these two phenomena. The language familiarity effect for voice perception is unique in that access to the perceptual dimensions necessary for native-like (expert) talker identification are specifically embedded in linguistic knowledge. That is, like face perception, there may be certain universal dimensions in voice perception that listeners can draw upon for the task of talker identification, such as fundamental frequency (pitch), timbre, or rate. (Although differences between languages may obfuscate these cues as well. Mandarin, for instance, is a tone language characterized by rapid pitch changes over individual syllables, whereas pitch contours in English vary over larger portions of speech.) Unlike face perception, however, many important individuating features of voices are only meaningful with regard to the phonological structure of one language versus another. For example, although the formant structure (vocal

tract resonance frequencies) of a vowel can provide information about the dimensions of the vocal tract, it is also crucial to know which vowel is being spoken in order to be able to encode the variation relevant to talker identification. Consider Mandarin, which has two vowels /u/ and /y/, both of which are generally perceived by English speakers as the English vowel /u/. Since Mandarin /y/ is distinguished from Mandarin/u/by the height of the second formant, a native English listener unaware of the phonemic contrast may perceive these two sounds as an /u/ produced by a speaker with either a short or long oral cavity, respectively. Extended exposure to these features in the absence of the linguistic knowledge necessary to make the variation meaningful is unlikely to improve individuals' ability to successfully use these dimensions in talker identification. We believe this is why our present study demonstrated that, unlike the face learning studies, the magnitude of the language familiarity effect remained constant throughout training when subjects have no familiarity in the foreign language (see Fig. 2A). Only in the case where subjects bring additional linguistic knowledge to bear (i.e., second-language learners, Fig. 2B) does the magnitude of the language familiarity effect decrease with training, similar to the decrease in magnitude of the own-race bias in face perception demonstrated by Malpass et al. (1973). We acknowledge that differences between speakers' and listeners' race may additionally contribute to impairment in talker identification ability. Because the target languages in the present study are predominately spoken by different races, it remains possible that part of the overall effect seen here may be due to physical differences in vocal tract anatomy between races. However, it bears noting that even though their results were not sufficient to answer the questions addressed by the face perception literature and our present study, Goggin et al. (1991) did demonstrate a similar effect when investigating two languages spoken predominately by the same race, English and German.

The integration of speech and voice perception for human listeners has broad societal implications in addition to understanding the structure of auditory cortex. For example, an important goal of electrical and computer engineering is developing efficient and accurate computer systems for talker identification. Such systems have desirable application for security and intelligence purposes. However, the accuracy of such systems remains far from optimal, and may benefit substantially from being able to take into account phonetic variation among talkers (e.g. Li & Espy-Wilson, 2004). Access to this phonetic information, either by computers or by human listeners, necessarily requires underlying linguistic knowledge. Our results also have considerable import in the field of forensic voice identification. We have shown that both utterance content and language can affect individuals' ability to identify talkers. Such results should give one pause when considering the testimony of "earwitnesses" who are identifying the voices of a speaker of a language they themselves do not know. Similarly, the selection of individuals for employment in fields where accurate voice identification is important, such as forensics and crime prevention or intelligence and national security, should bear in mind the importance of drawing on individuals with proficiency in the relevant languages.

Altogether, we have seen that higher-level cognitive auditory processes such as talker identification are unlikely to be subserved by singular brain regions, but instead involve multiple, integrated bilateral regions of auditory cortex otherwise responsible for processing both voice and speech. That these regions extensively interact in processes such as speech and voice perception should be taken into account during future studies of the brain bases of human auditory and linguistic capacity.

Acknowledgments

The authors wish to thank Cynthia Clopper, Ann Bradlow, Joan Chiao, Tasha Dees, Jay Mittal, Gnyan Patel, Janet Pierrehumbert, Matt Goldrick, and three anonymous reviewers for their assistance. This research was supported by Northwestern University and National Institutes of Health grants HD051827 and DC007468 to PW.

Appendix A

A.1. English training sentences

- The boy was there when the sun rose.
- A rod is used to catch pink salmon.
- The source of the huge river is the clear spring.
- Kick the ball straight and follow through.
- Help the woman get back to her feet.

A.2. English generalization sentences

- A pot of tea helps to pass the evening.
- Smoky fires lack flame and heat.
- The soft cushion broke the man's fall.
- The salt breeze came across from the sea.
- The girl at the booth sold 50 bonds.

A.3. Mandarin training sentences

- 院子门口不远处就是一个地铁站.
  yuan zi men kou bu yuan chu jiu shi yi ge di tie zhan
  "There is a subway station not far from the entrance to the yard."

- 这是一个美丽而神奇的景象.
  zhe shi yi ge mei li er shen qi de jing xiang
  "This is a beautiful and magical scene."

- 树上长满了又大又甜的桃子.
  shu shang zhang man le you da you tian de tao zi
  "The treetop has grown full of big, sweet peaches."

- 海豚和鲸鱼的表演是很好看的节目.
  hai tun he jing yu de biao yan shi hen hao kan de jie mu
  "The dolphin and whale performance is a very good program to watch."

- 邮局门前的人行道上有一个蓝色的邮箱.
  you ju men jian de ren xing dao shang you yi ge lan se de you xiang
  "On the sidewalk in front of the post office there is a blue mailbox."

A.4. Mandarin generalization sentences

- 天文望远镜可以用来观察天空.
  tian wen wang ya jing ke yi yong lai guan cha tian kong
  "An astronomical telescope can be used to observe the sky."

- 她到过很多地方观光旅游.
  ta dao guo hen duo di fang guang guang lu you
  "She has been many places to visit."

- 山间的小道蜿蜒曲折.
  shan jian de xiao dao wan yan qu zhe
  "The mountain path winds torturously."

- 春天来了, 山上开满了樱花.
  chun tian lai le, shan shang kai man le ying hua
  "Spring has come, the mountain top has bloomed with cherry blossoms."

- 下雪以后, 田野里白皑皑的一片.
  xia xue yi hou, tian ye li bai ai ai de yi pain
  "After the snow, the field was a snow-white patch."

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: University of Chicago Press.

Alain, C., Arnott, S. R., Hevenor, S., Graham, S., & Grady, C. L. (2001). 'What' and 'where' in the human auditory system. In *Proceedings of the National Academy of Science 98* (pp. 12301–12306).

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.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *3*, 129–135.

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

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

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–312.

Best, C. T. (1994). The emergence of native-language phonological influence in infants: A perceptual assimilation model. In J. Goodman & H. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge, MA: MIT Press.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*, *61*, 206–219.

Caldara, R., & Abdi, H. (2006). Simulating the other-race effect with autoassociative neural networks: Further evidence in favor of the face-space model. *Perception*, *35*, 659–670.

Clopper, C. G., & Pisoni, D. B. (2006). Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language Variation and Change*, *18*, 193–221.

Elliott, E. S., Wills, E. J., & Goldstein, A. G. (1973). The effects of discrimination training on the recognition of White and Oriental faces. *Bulletin of the Psychonomic Society*, *2*, 71–73.

Fecteau, S., Armony, J. L., Joanette, Y., & Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An FMRI study. *Neuroimage*, *23*, 840–848.

Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, *3*, 191–197.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory and Cognition*, *19*, 448–458.

Golby, A. J., Gabrieli, J. D. E., Chiao, J. Y., & Eberhardt, J. L. (2001). Differential responses in the fusiform gyrus to same versus other-race faces. *Nature Neuroscience*, *4*, 845–850.

Goldinger, S. D. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 33–66). San Diego, CA: Academic Press.

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage*, *21*, 494–506.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K., Kojima, S., & Nakamura, K. (1997). Vocal identification of speaker and emotion activates different brain regions. *NeuroReport*, *8*, 2809–2812.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization. *Journal of the Acoustical Society of America*, *88*, 642–654.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*, 759–763.

Kreiman, J., & Van Lancker, D. (1988). Hemispheric specialization for voice recognition: Evidence from dichotic listening. *Brain and Language*, *34*, 246–252.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*, 606–608.

Li, G. J., & Espy-Wilson, C. (2004). A novel dynamic acoustical model for speaker verification. In *Proceedings of the 147th meeting of the Acoustical Society of America*.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*, 358–368.

Licklider, J. C. R. (1952). On the process of speech perception. *Journal of the Acoustical Society of America*, *24*, 590–594.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own- and other-race faces. *Journal of Personality and Social Psychology*, *13*, 330–334.

Malpass, R. S., Lavigueur, H., & Weldon, D. E. (1973). Verbal and visual training in face recognition. *Perception and Psychophysics*, *14*, 283–292.

Meissner, C. A., & Brigman, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces. *Psychology, Public Policy, and Law*, *7*, 3–35.

Mullenix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception and Psychophysics*, *47*, 379–390.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., & Kojima, S. (2001). Neural substrates for recognition of familiar voices: A PET study. *Neuropsychologica*, *39*, 1047–1054.

Neuner, F., & Schweinberger, S. R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. *Brain and Cognition*, *44*, 342–366.

Nusbaum, H. C., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullenix (Eds.), *Talker*

variability in speech processing (pp. 109–132). San Diego, CA: Academic Press.

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

Open Speech Repository (2005). Open Speech Repository. Retrieved August 2005 from the World Wide Web: http://www.voiptroubleshooter.com/open_speech/index.html

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 309–328.

Pisoni, D. B., Saldana, H. M., & Sheffert, S. M. (1996). Multi-modal encoding of speech in memory: A first report. In *Proceedings of the international congress on spoken language processing* (pp. 1664–1667).

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 651–666.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology*, *32*, 97–127.

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1447–1469.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Slone, A. E., Brigham, J. C., & Meissner, C. A. (2000). Social and cognitive factors affecting the own-race bias in Whites. *Basic & Applied Social Psychology*, *22*, 71–84.

Small, S., & Nusbaum, H. C. (2004). On the neurobiological investigation of language in context. *Brain and Language*, *89*, 300–311.

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

Tarter, V. (1984). Laterality differences in speaker and consonant identification in dichotic listening. *Brain and Language*, *23*, 74–85.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*, 121–131.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race on face recognition. *Quarterly Journal of Experimental Psychology A*, *43*, 161–204.

Van Lancker, D., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*, 185–195.

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., & Giraud, A. L. (2004). Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage*, *22*, 948–955.

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

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Wong, P. C. M. (2002). Hemispheric specialization of linguistic pitch patterns. *Brain Research Bulletin*, *59*, 83–95.

Wong, P. C. M., Nusbaum, H. C., & Small, S. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*, 1173–1184.

Wong, P. C. M., Parsons, L. M., Martinez, M., & Diehl, R. L. (2004). The role of the insula cortex in pitch pattern perception: The effect of linguistic contexts. *Journal of Neuroscience*, *24*, 9153–9160.

Wong, P. C. M., & Perrachione, T. K. (in press). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistic*.

Wong, P. C. M., Perrachione, T. K., & Parrish, T. B. (in press). Neural characteristics of successful and less successful speech and word learning in adults. *Human Brain Mapping*.

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