

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**  
Progress Report No. 27 (2005)  
*Indiana University*

**Identification of Bilingual Talkers across Languages<sup>1</sup>**

**Stephen J. Winters, Susannah V. Levi and David B. Pisoni**

*Speech Research Laboratory  
Department of Psychological and Brain Sciences  
Indiana University  
Bloomington, Indiana 47405*

---

<sup>1</sup> This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant DC-00012 and NIH-NIDCD Research Grant R01 DC-00111). We would like to thank Christina Fonte, Jen Karpicke, and Melissa Troyer for their help in running subjects and editing stimuli.

## **Identification of Bilingual Talkers across Languages**

**Abstract.** Two groups of monolingual, native English-speaking listeners were trained to identify the voices of ten German-English bilingual talkers. One group of listeners learned to identify the voices from English stimuli only, while the other group learned to identify the talkers from German stimuli only. After four days of training, both groups of listeners were asked to identify the same talkers from novel stimuli in both the language they had been trained on and the language they had not heard during training. No differences were observed in the amount of improvement in talker identification accuracy made by the two groups of listeners during training. In testing generalization across languages, however, the English-trained listeners performed significantly worse on German stimuli than they did on English stimuli, while the German-trained listeners identified talkers just as well from English stimuli as they did from German stimuli. This pattern of generalization across languages suggests that some of the indexical properties of speech are language-specific, while others are language-independent. The English-trained listeners apparently learned to identify talkers by relying on language-specific indexical information, while the German-trained listeners learned to identify talkers through language-independent indexical information. This pattern of results suggests that listeners may follow a mandatory perceptual strategy whereby they make use of language-specific indexical information when they can understand the language that is being spoken; otherwise, they learn to identify voices on the basis of language-independent information alone. This perceptual tendency may result from the influence of automatic linguistic processing on listener performance in a talker identification task that requires conscious control.

### **Introduction**

Traditionally, linguists have distinguished between the linguistic and the indexical properties of speech (Abercrombie, 1967). Indexical properties of speech contain information about personal characteristics of the speaker—such as the speaker’s age, gender, sociolinguistic background or personal identity—while linguistic properties carry information about the message the speaker is trying to convey. While both indexical and linguistic information is simultaneously transmitted to listeners in the same speech signal, the extent to which these properties of speech may interact with each other—either in the speech signal itself or in the process of speech perception—has long been a matter of debate. There are two competing models of how these types of information are processed in the perception of speech: the modular view and the integrated view. In brief, the modular view assumes that the indexical properties of speech and the linguistic properties of speech are processed independently of one another and do not interact in speech perception. The integrated view, on the other hand, holds that indexical and linguistic properties are inextricably bound to one another in speech and that they affect each other in language processing and other tasks.

#### **Modular View**

In first characterizing the distinction between the linguistic and indexical properties of speech, Abercrombie (1967) described the indexical properties as “extra-linguistic,” and argued that information about the “medium” or the “source” of the message is not relevant to linguistic communication. This characterization implies that the perceptual process of recognizing a talker, or identifying some of that talker’s personal characteristics, can operate independently of the process of perceiving the linguistic content of an utterance. The listener simply has to identify which properties of the signal derive from the

talker and which derive from the linguistic system of phonological contrasts. Similarly, other researchers have assumed that speech perception involves a process of “talker normalization” (see Pisoni, 1997 for a review) which strips away the talker-specific information in speech and yields linguistic representations that are abstract and talker-independent (Halle, 1985). This “modular” view of speech perception holds that the process of identifying the linguistic content of a spoken utterance essentially involves identifying those linguistic properties of the signal which are independent of the talker.

There is clear evidence from both behavioral and neurological studies that the linguistic and indexical properties of speech can be processed independently of one another. For example, listeners can identify the linguistic content of spoken messages that are largely devoid of talker-specific information. Several studies have shown that listeners can identify talkers from time-reversed samples of speech, the linguistic content of which is unintelligible (Bricker & Pruzansky, 1968; Clarke, Becker, & Nixon, 1966; Williams, 1964). The same independence of talker and linguistic information has also been found, to a lesser extent, in filtered speech (Compton, 1963; Pollack, Pickett, & Sumbly, 1954) and whispered speech (Pollack, Pickett, & Sumbly, 1954; Williams, 1964). Phonagnosia, a phenomenon in which neurologically-impaired listeners can comprehend spoken utterances in a language that they know but cannot identify the voices of familiar talkers, also provide converging evidence that the linguistic processing of speech can take place independently of talker recognition (Van Lancker, Cummings, Kreiman, & Dobkin, 1988).

Other behavioral studies have shown that voice and linguistic information appear to be processed in different parts of the brain. In an early study of hemispheric specialization, Landis, Buttet, Assal, and Graves (1982) found that listeners utilize both hemispheres in voice recognition, whereas there was a distinct advantage of the left hemisphere for linguistic tasks (e.g., word recognition). In one experiment, Landis et al. played monosyllabic consonant-vowel words into either the right or the left ear, and asked listeners to press a button every time they heard a specific target word. The listeners’ reaction times showed a clear right-ear advantage (REA) for this linguistic task. In a second experiment, listeners were asked to push a button when they heard a particular male or female voice. For this task, no clear advantage for one ear over the other was found. Instead, the results revealed a REA when the target voice was female, but a left-ear advantage (LEA) when the target voice was male. Landis et al. interpreted these results by appealing to the fact that higher frequencies have been shown to elicit a REA and that female voices, with their higher fundamental frequency, may therefore also be processed with a REA. However, the stimuli used in the word recognition task were all presented in a female voice, so the REA found in that condition may have been due to the higher fundamental frequencies inherent to the stimuli, rather than a language-specific processing preference in the brain.

Kreiman and Van Lancker (1988) reported evidence of a dissociation between linguistic and indexical processing using a dichotic listening paradigm. In this paradigm, listeners heard different words played simultaneously in both ears. Each word was spoken in a different voice, selected from a database of fifty different famous male voices that the listeners knew. The listeners were asked to attend only to the stimulus in one ear or the other, and wrote down both the word that was played in that ear and the person who said the word. The listeners showed a clear REA in the word recognition task, but there was no significant advantage for either ear in the voice identification task.

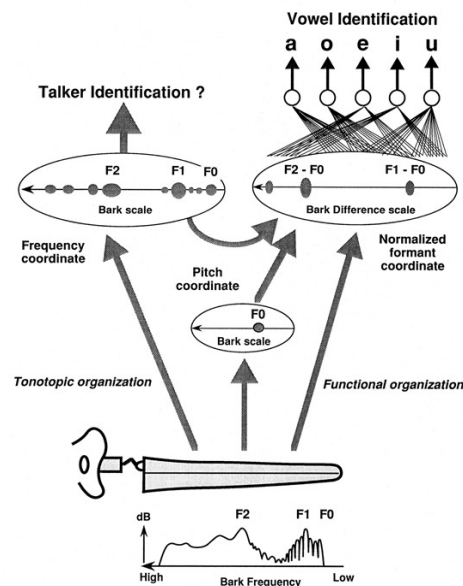
More recent studies have been able to isolate voice processing to more specific brain regions. Glisky, Polster, and Routhieaux (1995) tested elderly listeners' ability to recall either the content or the voice of previously heard sentences. They found that listeners with high frontal lobe function outperformed those with poor frontal lobe function on the voice task, but there were no differences between these two groups in their performance on the sentence content task. Conversely, listeners with

high medial temporal lobe function outperformed listeners with low medial temporal lobe function in the sentence content task, but there were no differences between these two groups of listeners on the voice task. More recently, using functional magnetic resonance imaging (fMRI), Stevens (2004) found that distinct brain regions were involved in voice- and word-discrimination tasks. Stevens presented pairs of words to listeners and asked them to either determine whether the talkers of the words were the same or whether the two words themselves were the same. Stevens found that the voice comparison task resulted primarily in activation in the right fronto-parietal area, whereas lexical processing was associated with the left frontal and bilateral parietal areas. These results indicate that, to some extent, the processing of voice information takes place independently of the processing of linguistic information, in a different part of the brain.

Taken together, these behavioral and neurological findings suggest that there is a double dissociation between linguistic comprehension and talker recognition: both processes can, in certain circumstances, operate independently of one another. Furthermore, when listeners are asked to attend to voice characteristics of a speaker, they appear to utilize different areas of the brain than when they focus on the linguistic content of a spoken message.

### Integrated View

Other researchers have proposed that the linguistic and indexical properties of speech are closely coupled, both in the speech signal and in the process of speech perception. Figure 1, reproduced from Hirahara and Kato (1992), illustrates how both sources of information are encoded in an integrated fashion in the speech signal. The spacing between adjacent formants provides information about the vowels a talker has produced, while the absolute values of the same formants provide information about the talker's voice. Global acoustic-phonetic properties of the speech signal, like the values of vowel formants, may therefore be considered both "linguistic" and "indexical."



**Figure 1.** Schematic of acoustic properties which specify both linguistic and indexical information in the speech signal (from Hirahara & Kato, 1992).

More importantly, several studies have shown that the linguistic and indexical properties of speech interact with each other in perception. This interaction is bidirectional in nature: indexical properties affect linguistic processing and linguistic knowledge affects the processing of indexical information.

**Indexical Information Affects Linguistic Processing.** The influence of indexical information on linguistic processing has been shown in a series of studies that have systematically varied the number and type of voices used to produce the stimuli for linguistic processing tasks. Varying voice information in this way typically results in worse performance on the linguistic processing tasks. Mullennix and Pisoni (1990) first showed this effect by asking listeners to categorize a set of stimuli that varied along two different perceptual dimensions in a speeded classification task. Listeners either had to decide whether a target word began with a "p" or a "b" (i.e., a linguistic distinction) or whether the word was spoken by a male or a female talker (i.e., an indexical distinction). In one condition, Mullennix and Pisoni presented both "p" and "b" words to the listeners in a variety of different voices. In a control condition, all words were presented to listeners in the voice of a single speaker. Mullennix and Pisoni found that reaction times for the linguistic classification task were slower when the stimuli were presented in several different voices than when the stimuli were presented in just one voice. This result indicated that indexical information was not strictly "extra-linguistic" or irrelevant to linguistic processing. Instead, listeners had to take talker-specific voice information into account when performing a phoneme classification task.

In a series of studies, Goldinger (1996) showed that listeners exhibit a same voice advantage when performing a word recognition memory task. Goldinger first asked listeners, in a study phase, to perform a word recognition task, in which they typed a word that they heard presented in noise. In a test phase, the listeners then heard a series of words and were asked to indicate whether or not they had heard that word before in the word recognition task. Half of the words in this test phase were repeated items, and, of these, half were presented in the same voice as they had been presented in the study phase. Goldinger found that listeners more accurately identified test items as being repeated from the study phase when they were presented in the same voice in which they had been presented initially than when they were presented in a different voice.

Other studies have also shown that the indexical and linguistic properties of speech encoded and stored together in representations of spoken words in memory. This results in a "same-voice advantage" effect, whereby spoken word tokens are processed more efficiently and accurately by listeners when they hear those words spoken in the same voice as they have heard in past experiences. For instance, Goldinger, Pisoni, and Logan (1991) found a same-voice advantage effect in a list recall task. In this study, listeners first heard a list of 10 words and were subsequently asked to recall the list. Goldinger et al. varied both the number of voices in which list words were presented and the rate at which stimuli were presented. They found that, at fast presentation rates, lists with multiple talkers were recalled less accurately than lists that were spoken by only a single talker.

Schacter and Church (1992) found a similar same-voice advantage effect in a stem completion task. They initially presented words to listeners in a variety of voices and asked the listeners to rate how pleasant each word token sounded. Later, listeners were presented with the first syllables of those words in noise and were asked to write down the first word that came to mind. Listeners responded more often with words they had heard in the initial phase of the experiment when the words were re-presented in the same voice as the initial presentation than when they were presented in a different voice.

In a continuous recognition memory experiment using spoken words, Palmeri, Goldinger, and Pisoni (1993) played long lists of words to listeners and asked them to determine whether each word was an "old" word (one that had been previously heard) or a "new" word (one that had not been previously heard). In order to assess the effects of voice on the processing of words, half of the old words were presented again in the same voice and half were presented again in a different voice. As in the previous studies, listeners responded more quickly and accurately when old words were repeated in the same voice.

Several studies have also shown that familiarity with a set of talkers' voices can facilitate the processing of the linguistic content of novel messages produced by those talkers. Nygaard, Sommers, and Pisoni (1994) explored how voice familiarity aids linguistic processing by first training listeners to identify ten previously unfamiliar talkers, from individual spoken words, over a period of ten days. After training, listeners were tested on their ability to identify novel words spoken in noise by either the talkers they had learned to identify or by a set of unknown talkers. Nygaard et al. found that the listeners identified a significantly higher percentage of the novel words correctly when they were spoken by familiar talkers. In a follow-up experiment, Nygaard and Pisoni (1998) showed that this advantage of talker familiarity applies not only to individual words, but to sentence-length utterances as well.

**Linguistic Knowledge Facilitates Indexical Processing.** Not only does knowledge of the indexical properties of speech affect language processing, but linguistic knowledge also affects the processing of indexical information. Several studies have shown that the inability to understand the linguistic content of speech hinders talker identification. Thompson (1987) had native English-speaking participants listen to a paragraph read in either English, Spanish, or Spanish-accented English by a target talker, and then asked the listeners to identify the target talker from among six different voices after a one-week delay. Thompson found that listeners could identify talkers best in the English language condition, followed by the Spanish-accented English condition, and worst in the Spanish language condition. Goggin, Thompson, Strube, and Simental (1991) followed up on this study by presenting Spanish and English stimuli to both monolingual English-speaking and monolingual Spanish-speaking participants in a similar testing paradigm. They found that both groups of listeners were poorer at identifying the voice of the target talker when they did not understand the language.

It has also been shown that the facilitatory effect that knowledge of a language has on the ability to identify talkers extends to a listener's second language, as well. Listeners who have studied a target language as a second language (L2) identify voices in that language better than listeners who have no knowledge of the language (Schiller & Köster, 1996; Köster & Schiller, 1997; Sullivan & Schlichting, 2000). In particular, Schiller and Köster (1996) showed that listeners with no knowledge of German were significantly worse at identifying a target talker, speaking in German, than both L2 listeners and native German listeners. Interestingly, Schiller and Köster found that the L2 and native German listeners did not differ from each other in talker identification accuracy. Sullivan and Schlichting (2000) further showed that the extent to which listeners are familiar with a second language does not affect their ability to identify talkers, so long as they have some knowledge of the language. They found that L2 learners of Swedish all performed significantly better than listeners with no knowledge of Swedish in a talker identification task, but that the amount of exposure the listeners had to the second language (ranging from first year learners to fourth year learners) did not affect their ability to identify Swedish talkers. Sullivan and Schlichting also reported, however, that L2 learners did not reach the same level of proficiency in identifying Swedish talkers as native Swedish listeners did in their earlier study (Schlichting & Sullivan, 1997), though no statistics were presented to corroborate this claim.

Schiller, Köster, and Duckworth (1997) have shown that the facilitatory effect of language knowledge on talker identification disappears when the linguistic content of the signal is eliminated, in reiterate speech. Schiller et al. had German speakers read a passage using only the syllable [ma] and then tested native German listeners, native English listeners, and L2 learners of German attempt to identify the speakers of those passages. They found that the native German listeners did not perform any better at this task than either the L2 learners or the native English listeners, implying that the advantage that native listeners have over non-native listeners in identifying talkers in a given language disappears once of the linguistic content of spoken utterances has been removed.

**Summary: Previous Research.** The studies reviewed in this section suggest that linguistic and indexical information are closely coupled in the processing of speech. Strong effects of voice were observed in tasks which, on the surface, do not appear to rely on indexical or voice properties—such as word recognition or phoneme discrimination. Familiarity with a talker’s voice was also found to facilitate a listener’s ability to process the linguistic content of speech. Likewise, listeners can process spoken utterances that they have heard before more efficiently and accurately when they are presented again in the same voice than when they are presented in a different voice. Furthermore, listeners can identify talkers’ voices more accurately when they know the language in which an utterance is spoken.

**Current Study.** The results of previous research showing that language knowledge facilitates the ability of listeners to identify talkers are confounded by the fact that all of these studies changed talkers between language conditions. Since both the linguistic and the indexical properties of the stimulus materials changed between language conditions in these studies, it is not clear whether the listeners’ diminished performance in the unfamiliar language condition was due to their lack of knowledge of the linguistic properties of the unknown language or their lack of knowledge of whatever language-specific indexical properties the unfamiliar language might have. It is also unknown whether listeners can identify familiar talkers who are speaking in an unfamiliar language. That is—are the indexical properties that listeners use to identify a familiar voice in one language the same properties of speech that can be used to identify that voice in another language?

In order to investigate these questions, the current study was designed to investigate the ability of listeners to identify bilingual talkers, while they were speaking in two different languages. Listeners were first trained to identify the voices of these bilingual talkers while they were speaking in one language, and then tested on their ability to identify the same talkers while they were speaking in the other language. Any potential change in talker identification accuracy between language conditions would thus be due to the change in language, rather than any change in the specific talkers producing the stimuli. By separating the contributions of language and talker to the spoken test materials in this way, the present experiment provides a much stronger test of the extent to which the linguistic and indexical properties of speech interact with each other in the process of talker identification.

The modular view holds that the indexical properties of speech are extra-linguistic, and do not vary from language to language. If this is the case, then the indexical and linguistic properties of speech should not interact with one another in the process of talker identification. The language that a talker is speaking should not affect the ability of listeners to identify that talker’s voice because that talker’s indexical contribution to speech will remain constant from one language to another. In this experiment, listeners should therefore be able to generalize all of their knowledge of the bilingual talkers’ voices across languages; they should be just as good at identifying voices in the language that they have been trained on as they are at identifying the same voices in a language they have not heard before.

On the other hand, the integrated view holds that the linguistic and indexical properties of speech are closely coupled and interact with one another in the process of talker identification. In this case, the properties of speech that listeners use to identify a talker's voice differ from one language to another. The language that a talker is speaking should therefore affect the ability of listeners to identify that talker's voice. If listeners rely on language-specific indexical properties when learning to identify a talker's voice, they should not be able to identify the same talker's voice as well in an unfamiliar language, which may exhibit a different set of language-specific indexical properties. It should therefore be difficult for listeners to generalize their knowledge of the bilingual talkers' voices completely across languages in the proposed experiment. Instead, the listeners should be able to identify talkers more accurately when they are speaking in the language that they have been trained on than when they are listening to the talkers in an unfamiliar language.

The integrated view of speech perception does not preclude the possibility that some indexical properties might be language-independent, or shared across languages. Thus, some of the listeners' knowledge of talkers' voices should generalize across languages; i.e., their ability to identify a known set of talkers in an unfamiliar language should be better than their ability to identify a set of unknown talkers in a familiar language. It is, of course, possible to take an even stronger view of the extent of integration between the linguistic and indexical properties of speech and propose that all indexical properties are specific to the language which is being spoken. If this is the case, then there should be no generalization of talker knowledge across languages in a talker identification experiment such as this one, since whatever listeners know about what a talker's voice sounds like in one language would not hold for that same talker's voice in a different language. There is, however, little existing evidence or rationale for this strong theoretical standpoint to suggest that such results might emerge from this experiment, but it is worth considering here as a benchmark.

## Methods

### Stimulus Materials

Twelve female and ten male German L1/English L2 speakers who were living in Bloomington, IN, were recorded in a sound-attenuated IAC booth at the Speech Research Laboratory at Indiana University. Productions were recorded using a SHURE SM98 head-mounted unidirectional (cardioid) condenser microphone with a smooth frequency response from 40 to 20,000 Hz. Productions were digitized into 16-bit stereo recordings via Tucker-Davis Technologies System II hardware at 22050 Hz and saved directly to an IBM-PC Pentium I computer. Each speaker produced a single repetition of 360 English words and 360 German words. Each word was of the form consonant-vowel-consonant (CVC) and was selected from the CELEX English and German databases (Baayen, Piepenbrock, & Gulikers, 1995). German was selected as the second language in the experiment because it not only had a sufficient number of CVC words—which had the same phonotactic structure as the English CVC words—but also because there were uniformly calculated frequency counts for both the English and German sets of words in the CELEX database. Speakers read each word as it was presented to them on a computer monitor. Before each presentation, an asterisk appeared on the screen for 500 ms, signaling to the speaker that the next trial was about to begin. This was followed by a blank screen for 500 ms. After this delay, a recording period began which lasted for 2000 ms. The target word was presented on the screen for the first 1500 ms of this recording period. After the conclusion of the recording period, the screen went blank for 1500 ms, and then an asterisk appeared again to signal the beginning of the next recording cycle. The presentation of production items was blocked by language, but all within-language items were randomized for each speaker. Items that were produced incorrectly or too quietly were noted and re-recorded in the same manner following each recording block. The total recording time for each language

block was approximately one hour for each speaker. Speakers were given the option of recording both sets of language items on either the same day or on two separate days. All speakers elected to record all stimuli in a single recording session. The recording session took approximately two hours, and speakers were paid \$10 an hour for their time.

This process yielded recordings which were uniformly 2000 ms long. Since the actual productions of the stimulus word in each recording were always shorter than 2000 ms, the silent portions in the recording before and after each production were removed by hand using Praat sound editing software. All edited tokens were then normalized to have a uniform RMS amplitude of 66.499 dB.

Words from both languages varied in frequency based on counts from the CELEX database. Words varying in frequency of occurrence were included in the stimulus materials because listeners can identify high frequency words more quickly, and from less acoustic information, than low frequency words (Grosjean, 1980). We expected listeners to pay more attention to the acoustic/phonetic details of the low frequency words, and therefore develop a more robust mental representation of the acoustic/phonetic characteristics of the various talkers' voices from these tokens. For the purpose of analysis, the English words were divided into three equal groups of varying frequency. The 120 lowest frequency words all had a CELEX frequency count of less than or equal to 96, while the 120 highest frequency words all had a frequency of greater than or equal to 586. The remaining 120 words thus all had frequency counts between 96 and 586. The frequency count of homophones (e.g., rite, write, right) was taken to be the frequency count of the most frequent homophone; this homophone was also the word that the speakers were presented with during the recording sessions.

Ten speakers were selected as the training voices, based on their native language background and perceived nativeness in English. Speakers with southern German (N= 2), Austrian (N=3) and Romanian German (N=1) dialects were excluded from the set of training voices, along with speakers with self-reported speech or hearing disorders (N=2), and one speaker who did not finish the recording session. Of the remaining speakers, only the five male and five female speakers who were, on average, rated as being the least accented talkers (Levi, Winters, & Pisoni, 2005) were used in the talker identification training study.

## **Listeners**

All listeners were native English-speaking students at Indiana University in Bloomington, Indiana. None reported any knowledge of the German language prior to participation in the study. None of the listeners had ever lived in Germany or had any German-speaking friends or family members. All were right-handed and reported no known speech or hearing impairments at the time of the study. Participants were paid \$75 for their participation in the study. A total of 54 listeners participated in the study. Half were trained on English language stimuli, and half were trained on German language stimuli.

The response data from only 40 of these listeners was included in the statistical analysis of the results. Two of the listeners in the English training condition and four listeners in the German training condition did not complete the experiment. The data from listeners who did not correctly identify at least 40% of the talkers in 4 or more evaluation phases during training were also excluded from analysis. We considered 40% correct identification accuracy to be a reasonable level of performance for establishing that listeners had learned the talkers' voices during training, since 30% correct was significantly better than chance performance in each evaluation phase (excluding cross-gender confusions). Four participants did not meet this criterion in the English language group and two did not meet this criterion in the German language group.

There were twenty-one listeners in both language conditions who both completed the experiment and met the criterion for learning during the evaluation phases. In the English training group, 10 of these listeners heard the English language stimuli in the first generalization testing phase, while 11 heard the German language stimuli first in generalization. The data from the last participant who heard the German stimuli first in generalization was excluded from the statistical analysis, in order to balance the numbers between generalization block groups. Similarly, in the German training condition, 11 of the remaining listeners heard the English language stimuli first in generalization testing, while the other 10 heard the German stimuli first in generalization. The data from the last participant who heard the English stimuli first in generalization was excluded from the statistical analysis.

## Procedure

Participants were trained and tested in a quiet room. During training, each participant wore Beyer Dynamic DT-100 headphones while sitting in a front of a PowerMac G4. All stimuli were presented to participants over the headphones via a customized SuperCard (version 4.1.1) stack, running on the PowerMac G4.

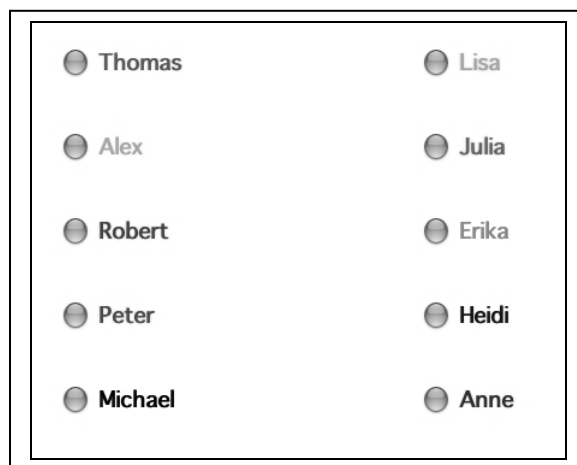
**Training.** Participants were trained to identify the ten different bilingual voices, by name, in eight training sessions spanning four days. The methodology used in these training sessions closely followed the methodology first developed by Nygaard, Sommers, and Pisoni (1994). Each training session consisted of seven distinct phases, which are summarized in Table 1.

Phase	Stimuli	Task
Familiarization #1	A set of five words, produced by all ten talkers	Listen and attend to voice/name pair
Re-familiarization #1	One word, produced by all ten talkers	Listen and attend to voice/name pair
Recognition #1	Sets of five different words for each talker, presented twice, in random order	Identify speaker of each word (feedback is provided)
Familiarization #2	A set of five words, produced by all ten talkers	Listen and attend to voice/name pair
Re-familiarization #2	One word, produced by all ten talkers	Listen and attend to voice/name pair
Recognition #2	Sets of five different words for each talker, presented twice, in random order	Identify speaker of each word (feedback is provided)
Evaluation	Sets of ten different words for each talker, presented once, in random order	Identify speaker of each word (no feedback provided)

**Table 1.** Summary of stimuli and tasks used during each phase in all training sessions.

In the familiarization phases, listeners heard a sequence of five words produced by each of the ten talkers. These words were the same for all ten talkers, and were presented one at a time. There was an inter-stimulus interval of 500 milliseconds between the presentation of each word. As each word was presented to the listener, the name of the talker who had produced the word was shown on the computer screen. Each talker had a name that was a common male or female name in both English and German. Each name was also presented in a unique and consistent color, in a unique and consistent position on the

screen. The layout of all ten names is shown in Figure 2. During this phase of training, participants did not respond to what they heard, but were instructed to pay attention to the names on the computer screen and listen to the sound of each talker's voice.



**Figure 2.** Layout of the ten talker names used in the experiment. (Names were in the following colors: Thomas, light blue; Alex, orange; Robert, red; Peter, purple; Michael, black; Lisa, green; Julia, dark pink; Erika, grey; Heidi, dark blue; Anne, brown).

After each familiarization phase, listeners underwent a brief re-familiarization phase in which they heard only one word spoken by all ten talkers. The same word was spoken by all ten talkers during re-familiarization. The participants did not register any response to the word but, again, were instructed to pay attention to both the name of the talker and the sound of the talker's voice.

In the recognition phases, listeners heard five different tokens, presented twice, from all ten talkers. Each word was presented in isolation to the listeners, whose task was to identify which talker had spoken each word. Participants identified talkers by clicking an on-screen button next to the appropriate talker's name (see Figure 1). After participants registered their responses, they received feedback, after a 333 millisecond interval, by hearing the stimulus token again, while the name of only the correct talker was presented to them on the computer screen. After receiving this feedback information, the listeners clicked an on-screen button to hear the next stimulus. The entire recognition phase was self-paced.

After the first recognition phase, listeners repeated the entire sequence of familiarization, re-familiarization, and recognition phases prior to beginning the evaluation phase. During the evaluation phase, listeners heard ten different words each from all ten talkers. As in the recognition phases, participants heard each word in isolation and were instructed to identify which talker had produced the word immediately after they heard it. Listeners did not, however, receive any feedback during the evaluation phase. Instead, they heard the next stimulus immediately after they registered their response to each stimulus.

The entire sequence of seven phases in each training sessions took most participants approximately 35 minutes to complete. Participants underwent two training sessions on each day of training, over the course of four days. Participants were required to take a short (approximately five minute) break between consecutive sessions on each day of training.

**Generalization.** After four days of training, all listeners participated in generalization testing on the fifth day of the experiment. Generalization testing began with two brief familiarization phases. In the first familiarization phase, listeners heard the same three words produced by all ten talkers. In the second, re-familiarization phase, the listeners heard the same word produced by all ten of the talkers. All of the words that were presented to the listeners in these familiarization phases were spoken in the same language that the listeners had heard during training. After the re-familiarization phase, the listeners were once again tested on their ability to identify the talkers from individual spoken words, in a series of two testing phases. The procedure used in these testing phases was identical to that used during the evaluation phase of each training session. Listeners heard one word at a time and were instructed to identify which talker had spoken the word. They received no feedback on their responses and were immediately presented with the next stimulus 500 milliseconds after registering their responses. The stimuli presented to the listeners in the two different generalization phases were in different languages. In one phase, the listeners heard words spoken in the language they had been trained on during the first four days of the experiment, while, in the other phase, they heard words spoken in the language they had not been trained on during the first four days of the experiment. Before testing, the listeners were instructed that the talkers might be speaking in an unfamiliar language. The order in which language blocks were presented in these two phases was counterbalanced across participants. For each participant, no more than two days intervened between any successive training days or the generalization test.

**Stimulus Selection.** The stimuli that were presented during training and generalization were independently selected for each listener from the larger set of individual word tokens in the bilingual talker database. For each listener, 100 words, balanced for frequency in each language, were first selected at random for use in the generalization testing blocks on the final day of the experiment. All 100 words that listeners heard in both generalization testing phases had thus never been presented before to the listeners during training. These 100 words consisted of ten different words spoken by all ten talkers, for both language blocks. No word, that is, was presented to listeners in more than one talker's voice during generalization.

After selecting out 100 words from the bilingual database for use in generalization, another 100 words were selected at random out of the remaining 260 items in the database, for each listener, for use in the familiarization and re-familiarization phases during training. These words were also balanced by frequency. Twelve of these items were presented to the listeners during each training session: five during the first familiarization phase, one during the first re-familiarization phase, five during the second familiarization phase, and one more during the second re-familiarization phase. Ninety-six of these words were thus presented to the listeners over the course of the eight training sessions, with the final four being presented to the listeners during the brief familiarization and re-familiarization phases prior to generalization testing (3 words and 1 word in these phases, respectively). No word that was presented during familiarization or re-familiarization was ever presented during generalization testing or in either the recognition or testing phases of the training sessions. The words selected for familiarization and re-familiarization were always in the same language as those words presented to the listener during the other phases of the training sessions.

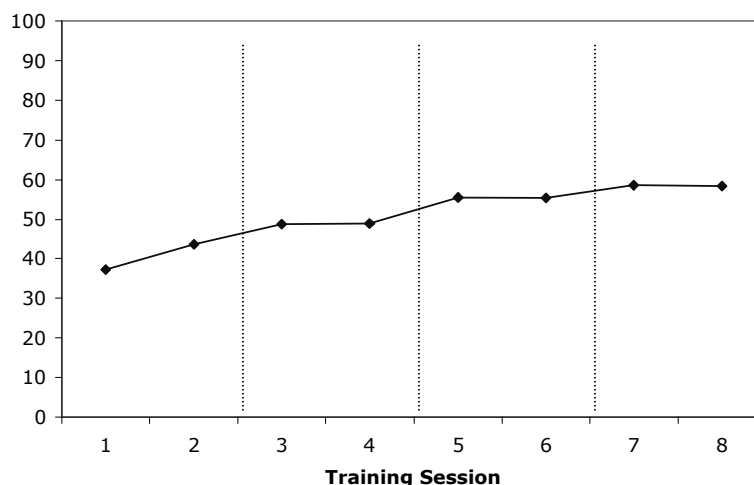
The remaining 160 words in the talker database were presented to each listener exclusively during the evaluation and recognition phases of the training sessions. These words were also balanced by frequency. For each training session, 20 words from this collection of 160 were selected at random for each talker for presentation to a particular listener. Five of these words were presented twice during the first recognition phase, while another five were presented twice during the second recognition phase. Talker-specific word tokens were presented more than once during these recognition phases because it

has been found that feedback does not facilitate perceptual learning unless the stimulus items that participants receive feedback on in a training paradigm are presented to them more than once (Winters, Levi & Pisoni, 2005). The remaining set of 10 items in each collection of 20 were then presented to the listeners, without repetition, in the evaluation phase of each training session. Over the course of the eight training sessions, then, listeners heard all 160 words as produced by all ten talkers. Within the evaluation and recognition phases of any given training session, however, listeners heard different sets of words produced by each talker. It was possible, therefore, for there to be overlap between the sets of words produced by each talker in any recognition or evaluation phase. In both the recognition and evaluation phases, all word tokens from all talkers were presented at random to the listeners, with the stipulation that no individual word was ever presented on consecutive trials.

## Results

### Training

A two-way, repeated measures Analysis of Variance (ANOVA) was run on the response data from the evaluation phases of the eight training sessions. This ANOVA investigated the effects that training session (1, 2, 3, 4, 5, 6, 7, 8) and training language (English, German) had on the percentage of talkers correctly identified in each testing phase. Training session was a within-subjects factor while training language was a between-subjects factor. The ANOVA revealed a significant main effect of training session ( $F(7,32) = 61.637$ ;  $p < .001$ ), but no effect (at the  $p < .05$  level) of training language, nor any interaction between training session and training language.



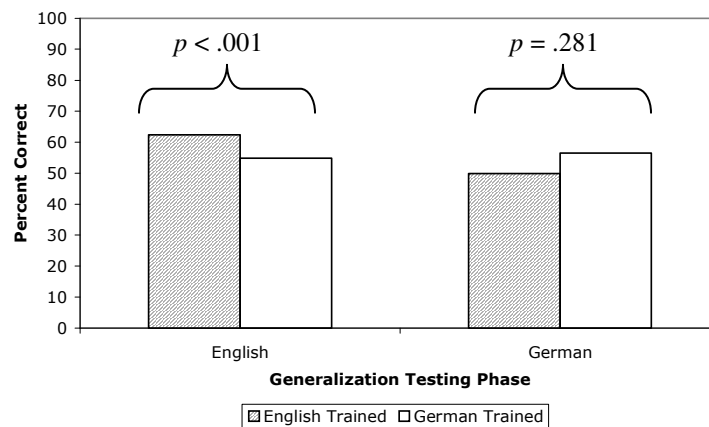
**Figure 3.** Percentage of talkers correctly identified, by all listeners, in the testing phase of each training session. Dotted lines denote breaks between separate days of the experiment.

Figure 3 shows the percentage of talkers that were correctly identified in the evaluation phases of each training session. Post-hoc, paired samples t-tests indicated that both groups of listeners consistently improved in identification accuracy over the duration of training. This improvement occurred in a step-wise fashion, however. Identification accuracy was significantly higher in training session two than in training session one ( $p < .001$ ). Accuracy was also significantly higher in training session three than in training session two ( $p = .002$ ). After session three, however, significant increases in identification

accuracy were only made between separate days of training. For instance, between sessions four and five—which occurred on days two and three of training, respectively—listeners’ average identification accuracy improved from 48.8% to 55.2% ( $p < .001$ ). Likewise, identification accuracy significantly improved between sessions 6 and 7 ( $p = .007$ ), which occurred across days three and four of training. Within a particular day of training, however, listeners did not significantly improve in identification accuracy between sessions ( $p > .825$ ).

## Generalization

A three-way, repeated measures Analysis of Variance (ANOVA) was run on the response data from just the generalization testing phases on the final day of the experiment. This ANOVA investigated the effects that testing language (English, German), training language (English, German), and generalization block order (trained language first, trained language second) had on the percentage of talkers correctly identified in each generalization testing phase. Testing language was a within-subjects factor while training language and generalization block order were between-subjects factors. The ANOVA revealed a significant main effect of testing language ( $F(1,36) = 27.687$ ;  $p < .001$ ), where accuracy was significantly better for English stimuli than German stimuli. There was also a significant interaction between testing language and training language ( $F(1,36) = 47.864$ ;  $p < .001$ ). All other main effects and interactions did not reach significance at the  $p = .05$  level.



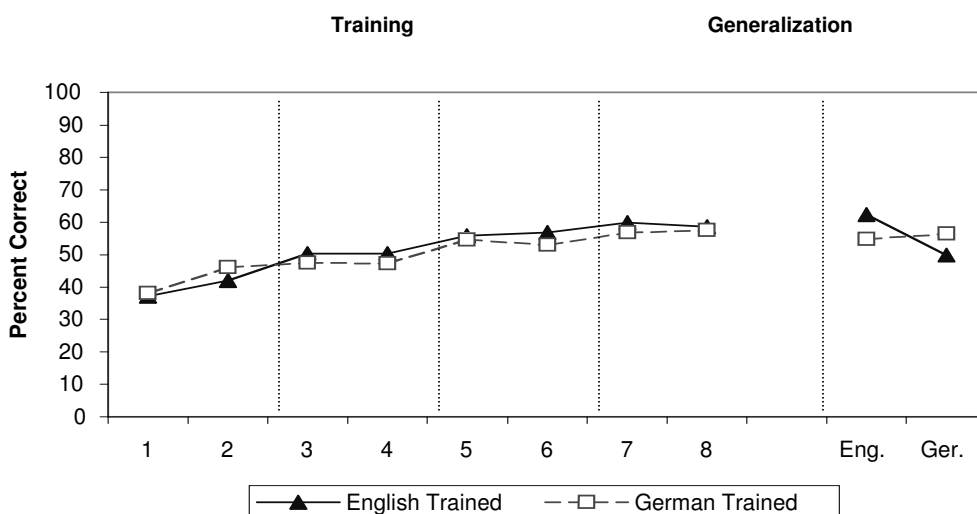
**Figure 4.** Percentage of talkers correctly identified, by each training group of listeners, in both generalization testing phases.

Figure 4 shows the percentage of talkers correctly identified, by each group of listeners, in the two generalization testing phases. Post-hoc analysis of the significant testing language by training language interaction indicated that the English-trained listeners demonstrated significantly higher talker identification accuracy on the English generalization block than on the German generalization block ( $p < .001$ ). The German-trained listeners, on the other hand, did not perform significantly better on the English generalization block than on the German generalization block ( $p = .281$ ). In comparing results across listener groups, post-hoc tests revealed that the German-trained group performed significantly better than the English-trained group on the German generalization block ( $p = .049$ ), while the English-trained group performed significantly better on the English generalization block ( $p = .016$ ).

## Combined Data

In order to assess the extent of generalization from training to novel stimuli, paired samples t-tests were conducted comparing the listeners' level of performance between each training session and the two generalization testing phases. Figure 5 shows the percentage of talkers correctly identified by each training group in both training and generalization.

For the English-trained listeners, there were no significant differences in talker identification accuracy between the English generalization block and the evaluation sessions on the final day of training ( $p = .095$  for session seven and  $p = .071$  for session eight). These listeners' performance on the English generalization block was, however, significantly better than their performance on the first six training sessions ( $p > .01$  in all cases). The English-trained listeners' performance on the German generalization block, on the other hand, was not significantly different from their performance on the third and fourth evaluation sessions, both of which took place on day two of training ( $p = .779$  and  $p = .826$ ). Their accuracy in German generalization was significantly better than their identification accuracy on day one of training ( $p < .01$ , for both sessions), but significantly worse than their identification accuracy on days three and four of training ( $p < .01$ , in all cases).



**Figure 5.** Percentage of talkers correctly identified, by each group of listeners, in the evaluation phase of each training session and in both generalization language blocks. Dotted lines denote breaks between separate days of the experiment.

Paired-samples t-tests also showed that the percentage of talkers correctly identified by the German-trained listeners in both generalization blocks was not significantly different than the percentage of talkers they correctly identified in training sessions five, seven and eight ( $p > .1$  in all cases). The German-trained listeners did identify a significantly higher percentage of talkers in German generalization than they did in training session six ( $p = .038$ ), but there was no significant difference between their performance in English generalization and in training session six ( $p > .1$ ). Otherwise, their performance in both generalization blocks was significantly better than in all evaluation phases on the first two days of training ( $p < .001$  in all cases).

## Lexical Frequency

The words that listeners heard during the evaluation phases were split into three frequency groups of equal size: 120 low frequency words (with frequencies ranging from 0-96), 120 mid-frequency words (97-588) and 120 high frequency words (greater than 588). The lexical frequency of the words presented during the evaluation phases of each training session did not significantly affect the ability of the listeners to identify the talkers who spoke them. The percentage of talkers that listeners identified correctly from low frequency words was 62.5%, while the corresponding percentages for the mid and high frequency groups of words were 59.4% and 64.2%, respectively. Paired samples t-tests revealed that none of these means were significantly different from one another ( $p > .08$  in all cases).

## Discussion

### Perceptual Learning during Training

The initial identification data from the training sessions showed that the training paradigm did, in fact, enable the listeners learn to identify the bilinguals' voices. Although a small minority of participants had some difficulty with the task, the majority of listeners significantly improved between the first and last training sessions in talker identification accuracy. The pattern of improvement in identification accuracy exhibited by these listeners was not consistent from one training session to the next, however. Listeners significantly improved in talker identification accuracy between the first and second training sessions, which were both on the same (first) day of training. After the second session, however, overall talker identification accuracy only improved significantly between sessions that took place on consecutive days. This pattern of learning suggests that some form of consolidation of what the listeners had learned took place in between consecutive training days, probably as a by-product of sleep (cf. Fenn, Nusbaum, & Margoliash, 2003). This pattern of improvement also suggests that listeners reached a learning plateau after the first training session on days two, three and four of the study, since they could no longer improve in their ability to identify talkers in the second training session on each of those days.

### Generalization: Effect of Training Language

The identification data from the generalization testing sessions showed that listeners could generalize their knowledge of the talkers' voices to novel stimuli in both languages. Both listener groups demonstrated that they could identify the bilingual talkers speaking in an untrained language at a significantly higher level than they could identify that same group of talkers, in the trained language, on the first day of training. For both groups of listeners, talker identification accuracy on novel stimuli in their trained language also did not decrease from their level of performance in the final training session. Listeners were thus able to generalize all of their knowledge of the talkers' voices to novel words within a language, and at least part of their knowledge of the talkers' voices across languages.

The extent to which the listeners could generalize their knowledge of the talkers' voices across languages depended on the language in which they had been trained to identify those voices. The listeners who had been trained on English stimuli significantly better talker identification accuracy from novel English words in generalization than they did from novel German words. With the novel English stimuli in generalization, these listeners performed just as well as they had on the English stimuli presented to them on the final day of training. With German stimuli, however, their identification accuracy decreased to a level equivalent to their performance on the second day of training (in sessions 3 and 4). Since the English-trained listeners' performance levels were still significantly better than their

accuracy level on the first day of training, their performance on the novel German words indicates that they were able to generalize some of what they had learned about the bilingual talkers' voices to a novel language. The German-trained listeners, on the other hand, showed complete generalization of talker knowledge across languages. These listeners performed just as well on the novel English stimuli in generalization as they had on the German stimuli presented to them in the final day of training. They also performed just as well on the novel German stimuli in generalization as they had on the final day of training. Thus, there was no decrease in performance when these listeners made the transition from training to novel stimuli in either language.

The inability of the English-trained listeners to generalize completely across languages suggests that some of the indexical properties that they used to identify talkers in training were language-specific. Since the German language lacks the language-specific indexical properties of English, the English-trained listeners were not able to use these properties to identify talkers when they heard the set of novel German stimuli in generalization. By the same token, the fact that the English-trained listeners were able to generalize some of their knowledge of the talkers' voices across languages suggests that they attended to and encoded some language-independent indexical properties during training, as well. Taken by itself, the pattern of generalization exhibited by the English-trained listeners thus conforms to the predictions made by the integrated model of speech perception: some indexical properties of speech are language-specific, while others are language-independent.

The generalization data from the German-trained listeners, however, appear to best fit the predictions made by the modular view of speech perception. The German-trained listeners showed complete generalization of their knowledge of the talkers' voices across languages, indicating that they had learned to identify the talkers' voices on the basis of language-independent indexical properties during training. Since this information—whatever it might consist of—does not change between English and German, the listeners did not suffer any decrease in identification performance when they were presented with novel stimuli in the English language in generalization testing.

### **Training Results: Non-effect of Training Language**

Although the language in which the listeners were trained affected how successful they were in generalizing to an untrained language, it did not affect the time course of perceptual learning during training itself; the amount of improvement in identification accuracy made by the participants during training in this study did not differ between the English-trained and German-trained groups of participants. The ability to understand the words that were presented to them in training did not, therefore, seem to provide the English-trained listeners with any additional advantage in the process of learning the talkers' voices. This finding conflicts with the results of earlier studies showing that it is easier for native English listeners to identify non-native talkers of English than talkers who are speaking a language other than English (e.g., Thompson, 1987; Goggin, Thompson, Strube, & Simental, 1991).

An effect of language on talker identification accuracy may not have emerged during the training portion of this study because its materials and methods were fundamentally different from those used in previous research. In voice line-up tasks, listeners are initially exposed to a short passage of speech from one talker and then asked to identify that talker, out of a variety of response options, on later trials. In the present study, listeners were familiarized with a larger set of talkers' voices before being tested on their ability to identify each talker from individual spoken words. The more involved training paradigm used in this study may therefore have reduced the native-language advantage that existed during the testing phase of previous voice identification studies. This advantage may also have been attenuated by the fact that only individual words were presented to listeners in this study. Nygaard and Pisoni (1998) have

shown that it is easier to identify talkers from sentence-long utterances than from individual word stimuli. Listeners can presumably use the higher-level semantic, syntactic and prosodic information in sentence- and paragraph-length utterances to identify talkers more easily when they can understand the language that is being spoken. With sentence stimuli, listeners also receive a longer sample of speech from a talker. Thus, listeners in the earlier voice line-up studies may have performed better on native-language stimuli because they had access to both sentence-length stimuli and higher-level linguistic information. Since the stimuli in this experiment lacked sentential and prosodic cues, however, the language in which the stimuli were spoken may have had less of an effect on talker identification accuracy during training. Finally, it is also possible that differences in talker identification accuracy during training were diminished because the same set of talkers was used in both language conditions. Although the finding of earlier research that talker identifiability is influenced by language has been replicated in several earlier studies, all of these studies consistently used different sets of talkers for different language conditions, and may therefore have confounded differences in the inherent distinctiveness of the voices with language-based difference in talker identification accuracy.

### **Training Results: Non-effect of Frequency**

The lexical frequency of the English words that listeners heard during training also did not affect their ability to identify talkers. The fact that the talker identification task did not require the listeners to access the lexicon may account for the absence of frequency effects on talker identification accuracy. Listeners could simply interpret each stimulus item in acoustic-phonetic terms without relying on higher-level lexical information to help them perform the task. The fact that the English-trained listeners did not need to access lexical information in order to perform the talker identification task may also account for their failure to perform at a higher level than the German-trained group, since they evidently never accessed linguistic information at a more abstract level than what the German-trained listeners could pick up from the acoustic/phonetic surface structure of the speech stimuli alone.

Such perceptual tendencies may actually have been helpful for training purposes, if paying more attention to the acoustic-phonetic properties of speech facilitates the learning of talker identity. Any effect that lexical access might have on a talker identification task could be tested in future research by requiring listeners to write down (or type) each word that they hear, before identifying the person who spoke it. Incorporating these additional processing operations into the talker-learning task could have a variety of effects on listeners' performance. The listeners might find it easier to identify talkers when they are speaking high frequency words, because this would make it easier for the listeners to access the lexical information necessary to do the word identification task. Conversely, listeners might do better when they are listening to low frequency words, because that would require them to pay more attention to the acoustic/phonetic details of the signal, and also require them to do more lexical processing before they are able to identify the word. Finally, the increased processing load may result in a stronger memory trace for both the low-frequency word and the talker who produced it (cf. Luce, Feustel, & Pisoni 1983).

### **Training Results: Poorer Performance than in Nygaard, Sommers, and Pisoni (1994)**

Listeners in this study also did not ultimately reach the same level of performance as the listeners reported in Nygaard, Sommers, and Pisoni (1994) did, even though the talker identification training paradigm in Nygaard et al. served as the basis for the one used in this study. Listeners in Nygaard et al. (1994) had to correctly identify 70% of the talkers on the final day of training in order to be included in the word identification transfer test on the final day of that experiment. About half of Nygaard et al.'s listeners were able to reach this criterion after nine days of training (18 out of 38). In the present study, however, only 10 of the 52 participants (six in the English-trained group, four in the German-trained

group) were able to correctly identify 70% of the talkers correctly in any evaluation session. The criterion for inclusion in this study was therefore reduced to 40% correct performance during testing in at least four different training sessions.

Our criterion was set lower than Nygaard et al.'s because it was not necessary, for the purposes of this study, to establish that the ability to identify talkers could facilitate a linguistic perception task such as word recognition in noise. We were only concerned with the extent to which listeners could generalize what they had learned about talkers' voices in training to novel stimuli, in different languages, on the last day of testing. All that was crucial to the success of this investigation, therefore, was that the listeners demonstrate that they were able to learn to identify the talkers' voices. Improvement to over 40% correct identification seemed to be a minimally satisfying demonstration of each listener's ability to have learned something about the various talkers' voices, since significantly better than chance performance in each training session was 30%. With this reduction in the criterion, only six out of 48 listeners (12.5%) failed to meet it after eight training sessions.

The poorer performance of the listeners in this study may be due to several methodological differences between this study's training paradigm and the one that was used in Nygaard et al.'s study (1994). Nygaard et al. trained listeners in nine sessions over nine separate days. Listeners in this study, however, participated in only eight training sessions, which took place over four days. The pattern of improvement during training in this study suggests that, after the first day of the experiment, it was necessary for listeners to sleep between training sessions in order for them to improve their performance. This result is consistent with the recent finding of Fenn, Nusbaum, and Margoliash (2003) that the perceptual learning of synthetic speech is enhanced by periods of sleep in between training sessions. For this reason, listeners did not show significant gains in talker identification accuracy between training sessions on the same day, after the first day of the experiment. The listeners may have been able to make such advances in identification accuracy between all eight training sessions, however, if those training sessions had all taken place on separate days. Spacing out the training cycles in this way could have enabled their performance to improve to the same level as that of the participants in Nygaard et al.

The listeners in Nygaard et al. (1994) also learned to identify the voices of native English talkers while, in this study, listeners learned to identify the voices of native German talkers who were speaking either English or German. Previous research by Goggin, Thompson, Strube, and Simental (1991) and Thompson (1987) has shown that English listeners have more difficulty identifying non-native speakers of English than native speakers of English. Hence, the native language of the speakers in this study may have contributed to the listeners' comparatively poorer level of performance in the talker identification task. However, some voices may also be simply more perceptually distinctive than others, regardless of their origin. It is thus possible that the voices of the talkers in Nygaard et al. just happened to be more distinctive than the ones used in this study, making the voice identification task easier for their listeners than it was for ours.

### **Generalization: Alternative Accounts**

Although the modular theory of speech perception accounts most gracefully for the German-trained group's generalization data, it is possible to construct an alternative account of this pattern of generalization in which the German-trained talkers learned to identify talkers using German-specific indexical properties. Since the generalization data from the English-trained group indicates that such language-specific information exists in English, similar language-specific information probably exists in the German language, as well. It is possible that the German-trained listeners in this study used such language-specific indexical information in learning to identify the talkers during training, in combination

with the same language-independent indexical information that was available to the English-trained listeners. The German-trained listeners may then have found it easier to generalize their knowledge of the talkers' voices to the English language stimuli because they were already familiar with the language-specific indexical properties that are unique to English (from native language experiences before the experiment). By combining this language-specific information with the language-independent indexical properties they had learned during training, the German-trained listeners could have identified the talkers' voices just as well in the English language generalization condition as they did in the German language condition.

The lack of an effect of language on talker identification accuracy during training argues against this interpretation of the generalization data, however. If the native English-speaking listeners did use German-specific indexical properties to identify talkers in the German language training condition, it should have taken these listeners some time to familiarize themselves with the novel indexical properties of the German language. There should, in other words, have been a gap in performance between the two training groups—at least for the first few training sessions—while the German-trained group learned how to make use of the German-specific indexical information. No such gap was observed in the training results, however, suggesting that the German-trained group used only language-independent information right from the beginning of the experiment to perform the talker identification task.

That the German-trained listeners might not have used German-specific indexical information to identify talkers is not surprising, because they had no knowledge or experience with the German language prior to the experiment. The English-trained listeners did know English before the experiment, however, and apparently relied extensively on what they knew about this language to help them perform the talker identification task in training. It is interesting to note, however, that this information evidently did not help the English-trained listeners perform any better in training than the German-trained listeners, who were using only language-independent information. The use of English-specific indexical information only affected the performance of the English-trained group by making it more difficult for them to generalize their knowledge of the talkers' voices to a novel language. As such, it is possible that the language-specific information English-trained listeners attended to during training did not actually help them perform the talker identification task. Instead, they may simply have been unable to ignore the irrelevant linguistic information in the signal—as long as they could understand it—possibly reflecting a failure of executive function and cognitive control (Schachar & Logan, 1990; Barkley, 1997).

Under this alternative interpretation, listeners engage in the linguistic processing of speech automatically—when they can understand the language that is being spoken—while talker identification is a non-automatic process that requires conscious attention and control. Mandatory linguistic processing may therefore affect the controlled process of talker identification in the same way that, for instance, the automatic process of reading words affects the controlled process of naming colors in the well-known Stroop Effect. Stroop (1935) had participants name the color in which different words were printed. Stroop found that participants named these colors more slowly when the word itself was the name of a different color than the ink in which it was printed. The information that the participants extracted from automatically processing the orthographic representations of the words thus interfered with the slower, controlled process of naming the color of the ink in which the word was printed. It has been shown that this interference effect is reduced, however, when the words are presented to participants in upside-down text, and therefore cannot be read them in an automatic fashion (Liu, 1973).

Analogously, linguistic information may have “interfered” with the process of talker identification in this experiment, when the listeners were presented with words in English and could therefore process them in an automatic fashion. Under these conditions, listeners may have based their

talker identification judgments on irrelevant linguistic information in the training stimuli. This linguistic information may therefore not be “integrated” with indexical information in the speech signal itself. Instead, the two sources of information may only become confused with one another during the process of speech perception. Interference between linguistic and indexical information would not occur when listeners cannot understand the linguistic content of the words automatically, as in the German language training condition. Without interference from linguistic information, the German-trained listeners would be able to process the indexical properties of speech in a more language-independent fashion than the English-trained listeners. The German-trained listeners’ representations of the talkers’ voices in memory would therefore be more robust and language-independent—and could generalize better across languages—than the English-trained listeners’ representations of the same voices. Similar effects of linguistic information interfering with indexical processing have recently been found in a same/different voice discrimination task (see Levi, Winters, & Pisoni, 2006).

### **Summary of Interpretation**

Participants in this study appear to be following a general perceptual strategy in which they make use of language-specific indexical information when it is available to them, regardless of what consequences that strategy might have for the generalizability of their perceptual representations for particular talkers. When listeners are identifying talkers who are speaking in a language they know, those listeners are able to process the indexical properties of speech in an integrated manner. When listeners are identifying talkers who are speaking in a language they do not know, however, those listeners process the indexical properties of speech in a modular, language-independent manner. Learning to identify voices in a modular fashion—on the basis of language-independent information only—makes it easier for listeners to generalize their knowledge of talkers’ voices to new languages. Relying on language-specific information to identify talker’s voices makes such generalization more difficult, but listeners do it anyway, when that information is available to them. Processing speech in an integrated manner, that is, apparently pre-empts the processing of speech in a modular fashion. Only when linguistic or indexical information is blocked in the speech signal—e.g., when listeners hear an unfamiliar language, are presented with filtered speech, or are suffering from phonagnosia—do listeners revert to a modular form of speech processing, which can operate without both forms of information in the speech signal.

### **Future Research**

By training German-English bilingual listeners in the same voice learning paradigm, it should be possible to test whether a decrease in talker identification accuracy across languages is due to a reliance on language-specific indexical properties or to an unfamiliarity with the language being generalized to. If all listeners automatically rely on language-specific indexical properties to identify talkers who are speaking a language they know, then bilingual listeners should rely on language-specific indexical properties in both the German and English language training conditions. These listeners should therefore have difficulty generalizing their knowledge of the talkers’ voices from one language to another, regardless of which language they have been trained in. If incomplete generalization across languages is the result of unfamiliarity with the language being generalized to, however, then bilingual listeners should exhibit no drop-off in talker identification accuracy in going from either English to German or from German to English in generalization, since they are familiar with both languages (and their attendant set of language-specific indexical properties).

Future research might also determine whether integrated linguistic and indexical information might facilitate performance across languages in either linguistic or indexical tasks, as well as hinder it. In this study, evidence for an interaction between the linguistic and indexical properties of speech came

from a significant decrease in performance by the English-trained listeners when they were tested on German stimuli in generalization. In this case, a reliance on language-specific information in training made the talker identification task more difficult when the listeners were required to generalize their knowledge of the talkers' voices to a different language. Past research, however, has indicated that the interaction between the linguistic and indexical properties of speech can also have facilitatory effects on linguistic tasks such as the recognition of words in noise. Nygaard, Sommers, and Pisoni (1994), for instance, found that listeners can identify novel words in noise better when they are spoken by talkers that those listeners have learned to identify, instead of talkers that those listeners have not heard before. This result has been taken as evidence that the linguistic and indexical properties of speech are not only integrated in perception, but that knowledge of language-specific indexical information is stored in memory and can facilitate the ability of listeners to carry out linguistic tasks.

Assuming a modular view of speech perception, however, it is possible that knowledge of the language-independent properties of a talker's voice might facilitate listeners' performance in a linguistic task. The more familiar listeners are with a particular talker's voice—in any given language—the easier it might be for them to filter out the indexical properties of a person's voice when attempting to identify the (talker-independent) linguistic properties of a word that person has spoken. It should be possible to test these alternative views of the relationship between the linguistic and indexical properties of speech in word recognition by training a group of listeners to identify the voices of German-English bilinguals from German stimuli only, and then testing those listeners on their ability to identify words in noise spoken by both the talkers they have learned to identify and an unfamiliar group of talkers. If language-independent knowledge of a talker's voice facilitates word recognition—as in the modular view—then listeners who have learned to identify a talker from German words only should be able to better identify English words spoken by talkers they have learned to identify. If only knowledge of language-specific indexical properties facilitates performance in a linguistic task, however, then the German-trained listeners should not improve in their ability to recognize English words spoken by either familiar or unfamiliar bilinguals. A study of this kind is currently under way in our laboratory.

## Conclusions

The present study investigated the extent to which the linguistic and indexical properties of speech are processed independently of one another by testing the ability of listeners to identify bilingual talkers' voices across two different languages. The extent to which listeners were able to generalize their knowledge of the bilinguals' voices from one language to another was considered within the context of two different views of speech perception. On the basis of the modular view of speech perception, which holds that linguistic and indexical information in the speech signal are processed independently of one another, in separate, perceptual channels, we predicted that listeners would be able to completely generalize their knowledge of the talkers' voices from one language to the other. However, on the basis of the integrated view of speech perception, which holds that the indexical properties of speech differ from language to language, we predicted that listeners would only show partial generalization of their knowledge of the talker's voices from one language to the other.

The results of this study suggest that listeners use both language-specific and language-independent indexical properties of speech. Listeners who were trained to identify bilinguals while they were speaking English showed incomplete generalization of their knowledge of the talkers' voices when they were asked to identify the same group of talkers while they were speaking German. In contrast, listeners who were trained to identify bilinguals while they were speaking German showed complete generalization of their knowledge of the talkers' voices when they were asked to identify the same group of talkers while they were speaking English. The English-trained group thus relied, in part, on indexical

properties that were specific to the English language in order to perform the voice identification task, while the German-trained group relied strictly on language-independent indexical information that could generalize across both languages.

Which features of a speaker's voice are language-independent, and which features are language-dependent? It may be assumed that the shape of a talker's vocal tract, nasal cavities and articulators have reliable effects on the acoustic output of that talker's speech, regardless of which language the talker is speaking. Rose (2003) points out, however, that the acoustic consequences of such "compulsory" features of a talker's voice may, in actuality, be very difficult for listeners to distinguish from one talker to another. Rose (2003) suggests, instead, that what makes talker's voices sound perceptually distinctive are the "chosen" features of their speech, which are under the talker's control to manipulate as he or she sees fit. In post-experiment debriefings, the participants in this study cited a number of different acoustic properties that they consciously listened for in attempting to identify each talker's voice. These features included qualities such as the pitch of the speaker's voice or the speed (i.e., the duration) with which a talker produced each word (e.g., some speakers consistently used a low pitch range or a high pitch range, while one female consistently produced each item with a very short duration). Such features of speech—while not necessarily "compulsory" aspects of a person's voice—could easily transfer from one language to another in bilingual talkers. A listener in a study such as this one could therefore identify a talker on the basis of perceiving such low-level acoustic qualities in either English or in a language with which they were not familiar, such as German. These "chosen" features of vocal identity might thus be considered "language-independent", so long as the languages that talkers are speaking do not require them to change acoustic characteristics such as pitch or duration in systematic ways (as in, for example, tone languages).

It is likely that listeners were also able to pick up on certain language-independent features of talkers' voices that were more complex than the basic acoustic properties of the speech signal. For instance, one listener, following the experiment, claimed that she could reliably identify one talker by the way she had "overexaggerated" the pronunciation of each word—in other words, by the fact that she had consistently hyperarticulated (Lindblom, 1990). Another listener claimed that she could consistently identify one of the male speakers by the fact that he sounded "gay." Such broad, phonetic features of a talker's voice may fall under the general rubric of a talker's "articulatory setting" (Rose, 2003). They could provide the listener with reliable, cross-linguistic cues to a talker's identity insofar as talkers are not required to change their articulatory settings by the phonetic rules of any given language.

On the other hand, phonetic markers of social identity (including sexual orientation, gender, class, regional affiliation, etc.) would be expected to change between languages—even two languages which are as phonetically similar as English and German. These phonetic attributes of the speech signal could thus serve as language-specific indexical properties. For instance, one phonetic marker of social identity which would almost certainly not transfer from one language to another is that of having a non-native accent in an L2. Many of the English-trained listeners cited the relative accentedness of each talker's speech as a feature they listened to in trying to identify talkers during training. Knowing how much of an "accent" a non-native talker has while they are speaking English is useless information to have when trying to identify the same talker when they are speaking German. The fact that many of the English-trained listeners claimed to have relied on perceived "accentedness" when identifying talkers in training may therefore account for the difficulty these listeners displayed in transferring their knowledge of the talkers' voices to novel German stimuli. (For a discussion of linguistic information on the perception of accentedness, see Levi, Winters, & Pisoni, 2005).

In the most general terms, the existence of both language-specific and language-independent indexical properties confirms the predictions made by the integrated view of speech perception. However, the results of this study suggest that listeners identify talkers on the basis of more than just language-independent or language-specific indexical properties. Another free parameter in the perceptual system appears to be the listener's strategy for doing the voice identification task. Listeners who understand the language that a talker is speaking will automatically make use of language-specific indexical properties to identify that talker's voice. They may even base their talker identification judgments on irrelevant linguistic information, if the automatic process of word recognition interferes with the controlled process of talker identification. If listeners cannot understand the language that a talker is speaking, however, they will be forced to identify that talker's voice on the basis of language-independent indexical information encoded in the speech waveform. Listeners can thus apparently switch between a modular form of speech perception and an integrated form of speech perception, depending on what information is available to them in the signal. The general perceptual strategy appears to be: make use of the most specific information which is available—including language-specific information—regardless of what consequences this might have for the construction of broadly generalizable perceptual categories for individual talkers.

The ability of listeners to make use of whatever information is available to them in the speech signal in order to perform linguistic and voice identification tasks demonstrates that the perception of speech is a highly robust and adaptive process. The fact that the perceptual system can rapidly adapt to changing listening conditions can also reconcile the apparently conflicting evidence for both the modular and integrated views of speech perception that was presented in the introduction. Speech perception operates in an integrated manner to the extent that listeners can and do use multiple sources of linguistic and indexical information in the speech signal to help them perform both linguistic and voice identification tasks more proficiently. When either linguistic or indexical information is removed from the speech signal, however, the perceptual system is capable of interpreting the linguistic or indexical information that is still available, in an independent and apparently modular fashion. The perception of speech may thus be either integrated or modular, depending on the context in which it operates. The evidence in favor of one view of speech perception does not necessarily invalidate evidence for the other, therefore, as long as the kind of information which is available to listeners in the speech signal is taken into account.

## References

- Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University.
- Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania [Distributor], Philadelphia, PA.
- Barkley, R.A. (1997). Behavioral inhibition, sustained attention, and executive functions constructing a unifying theory of ADHD. *Psychological Bulletin*, *121*, 65-94.
- Bricker, P.D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, *40*, 1441-1449.
- Clarke, F. R., Becker, R.W., & Nixon, J.C. (1966). Characteristics that determine speaker recognition. *Report ESD-TR-66-638*. Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.
- Compton, A.J. (1963). Effects of filtering and vocal duration upon the identification of speakers, aurally. *Journal of the Acoustical Society of America*, *53*, 1741-1743.
- Fenn, K.M., Nusbaum, H.C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature*, *425*, 614-616.

- Glisky, E.L., Polster, M.R., & Routhieaux, B.C. (April, 1995). Double dissociation between item and source memory. *Neuropsychology*, 9, 229-235.
- Goggin, J.P., Thompson, C.P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, 19 5, 448-458.
- Goldinger, S. D. (1996) Words and Voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166-1183.
- Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the locus of talker variability effects in recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 17, 152-162.
- Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, 28, 267-283.
- Halle, M. (1985). Speculations about the representations of words in memory. In V. Fromkin (Ed.), *Phonetic Linguistics*. (pp. 101-114). Academic Press: Orlando.
- Hirahara, T., & Kato, H. (1992). The effect of F0 on vowel identification. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure*. (pp. 89-112). Tokyo: Ohmsha Publishing.
- Köster, O., & Schiller, N.O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, 4, 18-28.
- Kreiman, J., & Van Lancker, D. (1988). Hemispheric specialization for voice recognition: Evidence from dichotic listening. *Brain and Language*, 34, 246-252.
- Landis, T., Buttet, J., Assal, G., and Graves, R. (1982). Dissociation of ear preference in monaural word and voice recognition. *Neuropsychology*, 20, 501-504.
- Levi, S.V., Winters, S.J., & Pisoni, D.B. (2005). Speaker-independent factors affecting the perception of foreign accent in a second language. In *Research on Speech Perception Progress Report No. 27* (pp. 49-64). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Levi, S.V., Winters, S.J., & Pisoni, D.B. (2006). Perception of the indexical properties of speech: universal or language-dependent? Poster presented at the *Tenth Conference on Laboratory Phonology*, Paris, France, June 30, 2006.
- Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & A. Marchal (Eds.), *Speech Production and Speech Modelling*. (pp. 403-439). Dordrecht: Kluwer
- Liu, A.-Y. (1973). Decrease in Stroop effect by reducing semantic interference. *Perceptual and Motor Skills*, 37, 263-265.
- Luce, P.A., Feustel, T.C., & Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25, 17-32.
- Mullennix, J.W., & Pisoni, D.B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, 47, 379-390.
- Nygaard, L.C., & Pisoni, D.B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60, 355-376.
- Nygaard, L.C., Sommers, M.S., & Pisoni, D.B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5, 42-46.
- 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. (1997). Some thoughts on “normalization” in speech perception. In K.A. Johnson and J.W. Mullennix (Eds.), *Talker Variability in Speech Processing*. (pp. 9-32). Academic Press: San Diego.
- Pollack, I., Pickett, J.M., & Sumbly, W.H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, 26, 403-406.

- Rose, P. (2002). *Forensic Speaker Identification*. London: Taylor & Francis.
- Schachar, R., & Logan, G.D. (1990). Impulsivity and inhibitory control in normal development and childhood psychopathology. *Developmental Psychology, 26*, 710-720.
- Schacter, D.L., & Church, B.A. (1992). Auditory priming: implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 915-930.
- Schiller, N.O., & Köster, O. (1996) Evaluation of a foreign speaker in forensic phonetic: a report. *Forensic Linguistics, 3*, 176-185.
- Schiller, N.O., Köster, O., & Duckworth, M. (1997). The effect of removing linguistic information upon identifying speakers of a foreign language. *Forensic Linguistics, 4*, 1-17.
- Stevens, A.A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research, 18*, 162-171.
- Stroop, J.R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.
- Sullivan, K.P.H., & Schlichting, F. (2000). Speaker discrimination in a foreign language: first language environment, second language learners. *Forensic Linguistics, 7*, 95-111.
- Thompson, C.P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1*, 121-131.
- Van Lancker, D.R., Cummings, J.L., Kreiman, J., & Dobkin, B.H. (1988). Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex, 24*, 195-209.
- Williams, C.E. (1964). The effects of selected factors on the aural identification of speakers. Section III, *Report ESD-TDR-65-153*. Hanscom Field, MA: Electronic Systems Division, Air Force Systems Command.
- Winters, S.J., Levi, S.V., & Pisoni, D.B. (2005). When and why feedback matters in the perceptual learning of the visual properties of speech. In *Research on Speech Perception Progress Report No. 27* (pp. 107-132). Bloomington, IN: Speech Research Laboratory, Indiana University.