Identification and discrimination of bilingual talkers across languages^{a)}

Stephen J. Winters,^{b)} Susannah V. Levi,^{c)} and David B. Pisoni^{d)}

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

(Received 9 May 2007; revised 2 April 2008; accepted 2 April 2008)

This study investigated the extent to which language familiarity affects the perception of the indexical properties of speech by testing listeners' identification and discrimination of bilingual talkers across two different languages. In one experiment, listeners were trained to identify bilingual talkers speaking in only one language and were then tested on their ability to identify the same talkers speaking in another language. In the second experiment, listeners discriminated between bilingual talkers across languages in an AX discrimination paradigm. The results of these experiments indicate that there is sufficient language-independent indexical information in speech for listeners to generalize knowledge of talkers' voices across languages and to successfully discriminate between bilingual talkers regardless of the language they are speaking. However, the results of these studies also revealed that listeners do not solely rely on language-independent information when performing these tasks. Listeners use language-dependent indexical cues to identify talkers who are speaking a familiar language. Moreover, the tendency to perceive two talkers as the "same" or "different" depends on whether the talkers are speaking in the same language. The combined results of these experiments thus suggest that indexical processing relies on both language-dependent and language-independent information in the speech signal.

© 2008 Acoustical Society of America. [DOI: 10.1121/1.2913046]

PACS number(s): 43.71.Bp, 43.71.Hw [PEI]

Pages: 4524–4538

I. INTRODUCTION

Speech researchers have traditionally distinguished between the *linguistic* and the *indexical* properties of speech (Abercrombie, 1967). Linguistic properties of speech provide information about the message that the speaker is trying to convey, while indexical properties provide cues to personal characteristics of the speaker—such as age, gender, sociolinguistic background, or emotional state. While both indexical and linguistic information are simultaneously transmitted to listeners in the same speech signal, it is an open question as to what extent these properties of speech may interact with each other in perception. One possibility is that listeners process the indexical and linguistic properties of speech independently of one another, while another possibility is that the indexical and linguistic properties of speech are inextricably bound to one another in the speech signal and therefore interact in both language processing linguistic and indexical tasks.

In this study, we investigated the extent to which linguistic and indexical properties interact in speech perception by testing the ability of listeners to identify and discriminate bilingual talkers across languages. We first trained listeners

to identify the voices of bilingual talkers from speech samples produced in one language only. We then tested their ability to identify or discriminate those same talkers' voices while they were speaking both the training language and a novel language. Since talker identification and discrimination are both indexical processing tasks, this study investigated the possible interaction of linguistic and indexical properties in speech perception by solely looking at the effects of language on indexical processing. If linguistic and indexical information do interact in speech perception, then performance in both tasks should depend on the particular language that the talkers are speaking. If, on the other hand, linguistic and indexical properties do not interact, then listeners' ability to identify and discriminate between talkers' voices should be independent of the language in which those talkers are speaking.

Paradoxically, the existing research literature provides evidence which suggests that the linguistic and indexical properties of speech interact in perception, and also that they may be independently processed. Evidence for the independent processing of linguistic and indexical information in speech comes from behavioral and neurological studies which indicate that listeners can successfully perform linguistic and indexical tasks when they do not have recourse to the other kind of information in the signal. For example, several studies have shown that listeners can identify talkers from time-reversed samples of speech, the linguistic content of which is unintelligible (Bricker and Pruzansky, 1966; Clarke *et al.*, 1966; Williams, 1964). The same independence of talker and linguistic information was also found to a lesser

^{a)} Portions of this work were presented at the 80th Annual Meeting of the Linguistic Society of America in Albuquerque, NM and LabPhon10 in Paris, France.

^{b)} Present address: Department of Linguistics, University of Calgary. Electronic mail: swinters@ucalgary.ca

^{c)} Present address: Department of Linguistics, University of Michigan. Electronic mail: svlevi@umich.edu

^{d)} Electronic mail: pisoni@indiana.edu

extent in filtered speech (Compton, 1963; Pollack *et al.*, 1954) and whispered speech (Pollack *et al.*, 1954; Williams, 1964). Phonagnosia, a phenomenon in which neurologically impaired listeners lose the ability to identify the voices of familiar talkers even though they can still comprehend spoken utterances in a familiar language, provides evidence that the linguistic processing of speech can also take place independently of talker recognition (Van Lancker *et al.*, 1988). Neurological research has also shown that indexical and linguistic information are processed in different parts of the brain. Landis *et al.* (1982) found hemispheric specialization for linguistic but not indexical information, while more recent studies isolated indexical processing to specific brain regions. Glisky *et al.* (1995) found that listeners with low frontal lobe function exhibited impaired indexical processing, while listeners with low medial temporal lobe function exhibited impaired linguistic processing. Additionally, Stevens (2004) found that voice discrimination primarily resulted in activation in the right frontoparietal area, whereas lexical discrimination was associated with the left frontal and bilateral parietal areas. Taken together, these behavioral and neurological findings suggest a double dissociation between linguistic comprehension and talker recognition: both processes can operate independently of one another.

In contrast, several studies have shown that the linguistic and indexical properties of speech do interact in perception. Furthermore, this interaction is bidirectional: indexical properties can affect linguistic processing, and linguistic knowledge can affect the processing of indexical information. The dependence of linguistic processing on indexical information has been documented in several studies that systematically varied the number and type of voices presented in linguistic processing tasks (Mullennix and Pisoni, 1990; Goldinger, 1996). Varying indexical information in this way consistently resulted in worse performance on these linguistic tasks. Other studies have also shown that the indexical and linguistic properties of speech are encoded and stored together in representations of spoken words in memory, thus facilitating the linguistic processing of messages spoken by familiar talkers (Goldinger *et al.*, 1991; Schacter and Church, 1992; Palmeri *et al.*, 1993; Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998).

The results of these studies were particularly influential in the development of exemplar-based theories of speech perception (Johnson, 1997; Goldinger, 1998; Pierrehumbert, 2001). These theories hold that listeners store individual experiences of speech—in a relatively unanalyzed form—in memory. Representations of linguistic categories thus consist of memory traces of particular words, spoken by particular talkers, in particular contexts, and at specific places and times. In these models, the categorization of new speech experiences operates on based on the combined, similarity-based activation response of all stored exemplars to incoming speech tokens. Since both indexical and linguistic information are stored together in the speech exemplars in memory, both of these properties may interact with each other in the processing of incoming speech tokens. In par-

ticular, speech produced either in a familiar language or by familiar talkers will be facilitated through activation of similar exemplars stored in memory.

A facilitating influence of linguistic knowledge on indexical processing was established by a variety of studies showing that talker identification is improved when listeners understand the language that is being spoken. For example, Thompson (1987) used a voice line-up task to test native English listeners' ability to identify talkers speaking in either English, Spanish, or Spanish-accented English. Thompson found that listeners could identify native English talkers best, followed by Spanish-accented English talkers, and, finally, Spanish talkers worst. Goggin *et al.*, (1991) followed up on this study by presenting Spanish and English stimuli to both monolingual English listeners and monolingual Spanish listeners in a similar testing paradigm. They found that both of the groups of listeners were poorer at identifying the voices of target talkers who were speaking an unfamiliar language.

This facilitatory effect of language familiarity on indexical processing extends to non-native languages (Schiller and Köster, 1996; Köster and Schiller, 1997; Sullivan and Schlichting, 2000). Schiller and Köster (1996) also found that native German listeners and non-native learners of German do not significantly differ in their ability to identify German talkers. Likewise, the extent to which listeners are familiar with a second language does not affect their ability to identify talkers as long as they have some knowledge of the language (Sullivan and Schlichting, 2000). The facilitatory effect of language knowledge on talker identification disappears, however, when the linguistic content of the speech is eliminated, as in reiterant speech (which replaces all syllables of a spoken message with the syllable [ma] but maintains the global prosodic patterns; see Schiller *et al.*, 1997).

One confound which is inherent to all of the studies that showed a facilitatory effect of language knowledge on talker identification accuracy, however, is that they consistently changed talkers between language conditions. It is therefore unclear whether the diminished performance of listeners in an unfamiliar language is due to the properties of the unfamiliar language itself or to the particular qualities of the talkers' voices that were presented in the unfamiliar language condition. The current study dissociates these effects by testing the ability of listeners to identify and discriminate the *same* group of talkers in two different languages. Any change that listeners exhibit in talker identification or discrimination accuracy between language conditions would thus be due to the change in language rather than to any change in the specific talkers producing the stimuli. By separating the contributions of the language and talker to the spoken test materials in this way, this experimental paradigm provides a stronger test of the extent to which the linguistic and indexical properties of speech interact in speech perception.

It is currently unknown whether listeners can generalize knowledge of talkers' voices across languages. Presumably, they could only do so if particular acoustic cues to individual talkers' voices are shared across languages. We will refer to such (potential) cues as *language-independent* cues to talker identity. The use of this term is not meant to imply that such cues are language universal; it is possible that a talker's

voice could be distinguished by cues that are shared across two phonologically similar languages (such as English and German) but would not necessarily be found in all languages that the talker is capable of speaking. Such cues might include, for instance, typical formant values for lax vowels, which are not often found in non-Germanic languages. Other more general language-independent cues to talker identity might include physical characteristics (such as the size and shape of the talker's vocal tract, nasal cavities, and vocal folds), age, or sex (Abercrombie, 1967). Nagao (2006) found that a talker's age can be reliably identified in both a known and an unknown language. Listeners were also shown to identify a talker's sex with a high degree of accuracy within a language, although it is unknown to what degree this ability carries over across languages (Lass *et al.*, 1976). The data we will present indicate that a talker's sex is identifiable even in an unknown language.

There are also potential language-dependent cues to talker identity. Abercrombie (1967) listed "group membership properties," such as regional or social markers, as indexical properties of speech. Such sociolinguistic markers might help identify a talker within one language but are unlikely to transfer over to another language. Other language-dependent indexical properties may overlap to some extent with a talker's physical characteristics. Todaka (1993), for instance, showed that Japanese-English bilinguals use different laryngeal settings in Japanese and English. Johnson (2005) also showed that gender-based properties of speech may change from language to language independently of a talker's sex and Nagao (2006) found that listeners can more accurately identify the age of talkers when they are speaking in a familiar language.

The extent to which knowledge of talkers' voices may generalize across languages depends not only on whether there is language-independent indexical information available to listeners in speech but also on whether listeners attend to that information when learning to identify voices. If listeners do identify voices by solely relying on language-independent information in speech, then the language that a talker is speaking should not affect voice identification accuracy. Listeners who identify talkers from these cues should be able to generalize voice knowledge without loss from one language to another. On the other hand, listeners who solely rely on language-dependent indexical cues to identify voices speaking in a particular language should not be able to generalize knowledge of those voices to a different language. Attending to such language-dependent cues to talker identity, however, should make it easier for those listeners to identify talkers in a familiar language than in an unfamiliar language.

These perceptual possibilities are not mutually exclusive. If indexical processing relies on both language-dependent and language-independent information in the speech signal, then some but not all of the listeners' knowledge of the talkers' voices should generalize across languages. In this case, the listeners' ability to identify a known set of talkers in an unfamiliar language would be better than their ability to identify a novel set of talkers in a familiar language. However, the same listeners should be more accu-

rate when identifying known talkers in a familiar language than in an unfamiliar language.

II. EXPERIMENT 1: BILINGUAL TALKER IDENTIFICATION

A. Methods

1. 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 high quality recording equipment and immediately digitized into 16 bit stereo recordings via a Tucker-Davis Technologies System II hardware at 22 050 Hz and directly saved to an IBM-PC Pentium I computer. Recordings were made of each speaker producing 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 *et al.*, 1995). German was selected as the second language in the experiment because it had a sufficient number of CVC words with the same phonotactic structure as the English CVC words and also because uniformly calculated frequency counts for both the English and German words were available in the CELEX database.

During recording, speakers read one-word prompts off of a computer screen while sitting in the sound-attenuated booth. These words were presented to speakers in randomized order and blocked by language. Any words that speakers produced incorrectly or too quietly were noted and rerecorded in the same manner following each recording block. An automated recording process yielded sound files that were 2000 ms long for each word. The silent portions in these sound files were later removed by hand using PRAAT sound editing software, and the resulting tokens were normalized to have a uniform rms amplitude of 66.499 dB. The total recording time for each language block was approximately 1 h for each speaker. All speakers recorded both language blocks in a single session and were paid \$10/h for their time.

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, words in both language blocks were divided into three equally sized groups of varying frequency (high, mid, low).

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

TABLE I. Summary of stimuli and tasks used during each phase in all of the training sessions in experiment 1.

Training session			
	Phase	Stimuli	Task
Training block I	Familiarization	Same five words produced by all ten talkers (500 ms ISI)	Listen and attend to voice/name pair
	Refamiliarization	Same one word produced by all ten talkers	Listen and attend to voice/name pair
	Recognition	Sets of five different words produced by each talker, presented twice in random order	Identify speaker of each word (feedback provided)
Training block II	Familiarization	Same procedure as above	Same procedure as above
	Refamiliarization	Same procedure as above	Same procedure as above
	Recognition	Same procedure as above	Same procedure as above
	Evaluation	Sets of ten different words produced by each talker, presented once in random order	Identify speaker of each word (no feedback provided)

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 rated as having the least foreign accent were used in the talker identification task (Levi *et al.*, 2007b).

Accent ratings for each talker were taken from a previous study in which individual words, which were spoken by the various talkers in the bilingual database, were rated on a Likert scale from 0 (“no foreign accent”) to 6 (“most foreign accent”) by native English listeners who had no familiarity with the German language (for more details, see Levi *et al.*, 2007b). Although we only included the bilingual talkers with the lowest accent ratings in experiments 1 and 2, these talkers were not “accentless.” The average, z -score-transformed accent ratings for the female talkers used in this study ranged from -0.27 to 0.22 , while the corresponding scores for the male talkers ranged from 0.02 to 0.69 —with the higher scores indicating a higher perceived degree of foreign accent. These talkers made few phoneme substitution errors in their word productions (most commonly coda voicing substitutions, such as “news” for “noose”), so these accent ratings probably reflected more subtle phonetic distinctions in their speech or perhaps lack of nativelike articulatory speed and fluency. Native English speakers who were rated in the same study were also not rated entirely accentless; they earned a z -score range of accentedness from -0.52 to -0.09 .

2. Listeners

All listeners were native English-speaking students at Indiana University in Bloomington, IN. None reported any knowledge of German prior 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. A total of 54 listeners partici-

pated in the study and were paid \$10/h for their participation. Half of these listeners were trained on English stimuli, and half were trained on German stimuli.

The response data from only 40 of these listeners were 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 the listeners who did not correctly identify at least 40% of the talkers in four 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 the 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. The last listener to complete the experiment in each of the two training conditions was also excluded from the statistical analysis, resulting in 20 listeners per group.

3. Procedure

Participants were trained and tested in a quiet room. All stimuli were presented to participants over Beyer Dynamic DT-100 headphones by a customized SuperCard (version 4.1.1) stack, running on a PowerMac G4.

(a) *Training.* Participants were trained to identify the ten different bilingual voices by name in eight training sessions spanning 4 days. The methodology used in these training sessions closely followed the methodology developed by Nygaard *et al.* (1994). The individual training sessions consisted of seven distinct phases, which are summarized in Table I.

In the familiarization phase, listeners heard the same sequence of five words produced by each of the ten talkers,

with an interstimulus interval (ISI) of 500 ms. As each word was presented to the listener, the name of the talker who had produced that word was shown on the computer screen. Each talker's name was a common male or female name in both English and German and was presented in a unique and consistent color in a unique and consistent position on the screen. During this phase of the 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. The refamiliarization phase was identical to the familiarization phase except that listeners heard only one word produced by each of the ten talkers.

After familiarization, listeners completed a recognition task in which they heard five different words, presented twice, from all of the ten talkers.¹ These stimuli were presented in a different random order for each participant. After the presentation of each word, listeners identified the talker of that word by clicking an on-screen button next to the appropriate talker's name. After the participants registered their responses, they received feedback by hearing the stimulus token again while the name of only the correct talker appeared on the computer screen. This portion of training was self-paced.

After completing two blocks of the familiarization, refamiliarization, and recognition phases, listeners completed an evaluation task. This evaluation phase was identical to the recognition task except that listeners did not receive feedback on their responses, and they heard ten different words from each talker, without any repetitions of the same word token. Each training session (consisting of two training blocks plus the evaluation phase) took approximately 30 min to complete. The participants completed two training sessions per day for 4 days and were required to take a short (approximately 5 min) break between consecutive sessions on each day of the training. For each participant, no more than 2 days intervened between any successive training days or the generalization test.

(b) *Generalization.* After eight training sessions, all listeners completed a generalization test on the final day of the experiment. The generalization testing began with a shortened familiarization phase, in which the listeners heard the same three words produced by all of the ten talkers followed by a refamiliarization phase. All of the words that were presented to listeners in these familiarization phases were spoken in the same language that listeners had heard during training. After familiarization, listeners identified talkers in two testing phases. These testing phases were identical to the evaluation phase at the end of each training session except that the stimuli were presented in different languages in each phase. In one phase, the listeners heard novel words spoken in the same language they had been trained on while in the other phase, they heard words spoken in the language they had *not* been trained on. Before the generalization, the listeners were informed that the talkers might be speaking in an unfamiliar language. The order in which trained and untrained language blocks were presented in these two generalization phases was counterbalanced across participants.

4. Stimulus selection

The stimuli presented during the 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 lexical frequency in each language—were randomly selected for use in the generalization blocks. These 100 words consisted of ten different words spoken by each of the ten talkers for both language blocks. Of the remaining 260 words in the database, 100—counterbalanced for lexical frequency—were randomly selected for use in the familiarization phases during the training. The remaining 160 words in the talker database were exclusively used during the evaluation and recognition phases of the training. In both the recognition and evaluation phases, all stimuli were presented in random order to the listeners. While different sets of words were selected for each talker in these phases, it was possible for there to be an overlap between the sets of words produced by each talker in the recognition phases. No individual word was ever presented twice on consecutive trials in recognition or evaluation testing.

B. Results

1. Training

A two-way, repeated measure analysis of variance (ANOVA) was conducted on the response data from the evaluation phases of the eight training sessions. This ANOVA investigated the effects that training session (1–8)—a within-subject factor—and training language (English, German)—a between-subject factor—had on the percentage of talkers correctly identified in each testing phase. The ANOVA revealed a significant main effect of training session [$F(7, 32)=61.637$; $p<0.001$] but no effect (at the $p<0.05$ level) of training language nor any interaction between the training session and training language.

Figure 1 shows the percentage of talkers that were correctly identified in the evaluation phases of each training session. *Post hoc*, paired sample *t* tests indicated that both of the groups of listeners consistently improved in identification accuracy over the duration of the training. This improvement occurred in a stepwise fashion. Identification accuracy was significantly higher in training session 2 than in training session 1 ($p<0.001$). Accuracy was also significantly higher in training session 3 than in training session 2 ($p=0.002$). After session 3, significant increases in identification accuracy only occurred between separate days of training [i.e., between sessions 4 and 5 ($p<0.001$) and between sessions 6 and 7 ($p=0.007$)]. Interestingly, this pattern of learning suggests that consolidation of learning took place only during sleep after the first day of training (Fenn *et al.*, 2003). More generally, these results indicate that the listeners were able to identify the voices of the bilingual talkers, and that all of the listeners learned to identify the talkers at the same rate regardless of the language in which they were trained.

2. Generalization

A three-way, repeated measure ANOVA was run on the response data from the generalization blocks on the final day

FIG. 1. Box plot of the percentage of talkers correctly identified by each group of listeners in the evaluation phase of each training session and in both generalization language blocks in experiment 1. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

of the experiment, with testing language (English, German) as a within-subject variable and both training language (English, German) and block order (training language first, training language second) as between-subject variables. This ANOVA revealed a significant main effect of testing language [$F(1,36)=27.687$; $p<0.001$], where accuracy was significantly better for English stimuli (58.6%) than that for German stimuli (53.2%). There was also a significant interaction between the testing language and training language [$F(1,36)=47.864$; $p<0.001$]. No other main effects or interactions were significant.

Figure 1 also 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 ($p<0.001$). The German-trained listeners, on the other hand, did not perform significantly differently on the English and German generalization blocks ($p=0.0281$). In comparing results across the listener groups, *post hoc* tests revealed that the German-trained group performed better than the English-trained group on the German generalization block,² while the English-trained group performed significantly better on the English generalization block ($p=0.016$).

3. Combined data

In order to assess the extent of generalization from training to novel stimuli, paired sample *t* tests were conducted by comparing the listeners' level of performance between each training session and the two generalization testing phases.

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 ($p=0.095$ for session 7 and $p=0.071$ for session 8). These listeners' performance on the English generalization block was, however, significantly better than their performance on the first six training sessions ($p>0.01$ in all cases). In contrast, their performance on the German generalization block was not significantly different from their performance on the third and fourth evaluation sessions, both of which took place on day 2 of the training ($p=0.779$ and $p=0.826$). Their accuracy in identifying talkers from novel German stimuli was significantly better than their identification accuracy on day 1 of the training ($p<0.01$ for both sessions) but significantly worse than their identification accuracy on days 3 and 4 of the training ($p<0.01$ in all cases). This pattern of results indicates that the English-trained listeners were able to successfully generalize some of their knowledge of the talkers' voices to German since their ability to identify the speakers of novel German tokens was significantly better than their performance on the English tokens in the first training session. These data also show that this generalization to German language stimuli was not complete; the listeners' performance in generalization was significantly better for novel English stimuli than that for novel German stimuli.

The paired-sample *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 5–8.³ However, their performance in both generalization blocks was significantly better than that

in all of the evaluation phases on the first two days of training (all $p < 0.001$). Thus, unlike the English-trained listeners, the German-trained listeners were able to generalize their knowledge of the talkers' voices—without a significant loss in identification accuracy—to a language they had not heard in training.

The words that the English-trained listeners heard during the evaluation phases of each training session were evenly split into three groups based on lexical frequency. The paired sample t tests revealed that the lexical frequency of the words presented during the evaluation phases of each training session did not significantly affect the listeners' ability to identify the talkers ($p > 0.08$ in all cases). The percentage of talkers correctly identified from low frequency words was 62.5%, for the mid-frequency, it was 59.4%, and for the high frequency, it was 64.2%.

C. Discussion

The results of this experiment show that there is sufficient language-independent information in speech to make the identification of familiar talkers across languages possible. Listeners steadily improved in their ability to identify talkers in both languages. Improvement in talker identification accuracy largely manifested itself between training days and did not significantly differ between the two training groups. In generalization, both of the groups of listeners identified familiar talkers better in the untrained language than they had identified those same talkers on the first day of the training. The gains in identification accuracy that listeners made in training thus carried over, in part, to stimuli produced by those same talkers speaking in a different language.

The extent to which the two groups of listeners could generalize their knowledge of the talkers' voices across languages depended, however, on the language in which they had been trained. The German-trained listeners exhibited a complete generalization of talker knowledge across languages—identifying talkers in English as well as they had identified talkers in German. The English-trained listeners, on the other hand, showed incomplete generalization of talker knowledge across languages, identifying talkers from novel English stimuli significantly better than they identified talkers from novel German stimuli. This pattern of results suggests that the two groups of the listeners processed the indexical cues in the speech tokens in different ways. The German-trained listeners apparently identified talkers by relying on language-independent indexical information in the signal. The representations of the talkers' voices that they developed in training therefore consisted of language-independent information that could be applied, without loss, to the identification of the same talkers speaking in a different language. The English-trained listeners, on the other hand, evidently learned to identify the bilingual talkers from language-dependent cues since they could not identify talkers in German as well as they had identified talkers in English.

The results of this experiment thus provide evidence for both language-dependent and language-independent indexi-

cal processings. Interestingly, listeners appeared to process indexical information in a language-dependent way when they could understand the language that was spoken; otherwise, they identified talkers' voices from only language-independent indexical information in the signal. This pattern of perceptual tendencies may provide insight into the apparently conflicting evidence for both views of speech perception in the existing research literature. The basic pattern in both previous research and this study seems to be that language-independent processing takes place only when the signal lacks linguistic information in some way (filtered speech, talkers speaking in an unfamiliar language, etc.) or when listeners are not capable of processing both kinds of information (phonagnosic listeners). When normal-hearing listeners receive an undegraded signal in a language that they can understand, however, they make use of all the information that is available to them, and the linguistic and indexical properties then interact. For this reason, *language-independent indexical processing* was exhibited by the German-trained listeners in this study, who learned to identify talkers speaking in a language they could not understand, while the English-trained listeners exhibited *language-dependent indexical processing* by relying on English-specific indexical cues that did not generalize to German.

Previous evidence for language-dependent indexical processing has been positive in nature, showing that listeners identify talkers better when they are speaking in familiar languages. Language-dependent indexical processing provided the English-trained listeners in this study with some processing benefits as well in both training and generalization. In generalization, the English-trained listeners not only performed significantly better on English stimuli than on German stimuli but they also performed better on the English stimuli than their German counterparts. That is, even though the German-trained listeners showed a "complete transfer" of their knowledge of the talkers' voices to English-language stimuli, they still could not identify talkers speaking in English as well as the English-trained listeners, and there were also subtle differences between the two listener groups in the patterns of improvement made during the training. There were no significant differences in identification accuracy between the two training groups for any given training session; however, the English-trained listeners performed better on the English stimuli in generalization than they had on all of the training sessions on the first three days of the training. In contrast, the German-trained group did not perform significantly better in generalization than they had on either the fifth or sixth training sessions (on the third day of the training). Thus, the English-trained listeners were somewhat more successful than the German-trained listeners at attaining an increasingly higher level of talker identification accuracy when generalizing to novel words spoken in the language presented during training.

The results of this study show that language-dependent indexical processing has negative implications as well. Relying on language-dependent information to identify talkers' voices in one language made it more difficult for the English-trained listeners to generalize their knowledge of the talkers' voices to a new language. Interestingly, this decrease

in talker identification accuracy occurred even though there is no a priori reason to believe that the English-trained listeners could not identify talkers from strictly language-independent indexical cues. However, the perceptual integration of linguistic and indexical information may automatically occur when listeners can understand the language that is being spoken. If so, such an automatic perceptual process offers only small gains in indexical processing within a known language at the expense of developing representations of talkers' voices which are more perceptually robust and generalizable to new languages.

III. EXPERIMENT 2: CROSS-LANGUAGE VOICE DISCRIMINATION

Experiment 2 further investigated the influence of language on indexical processing by testing the listeners' ability to discriminate voices both within and across languages. For this task, the listeners were asked to judge whether two speech stimuli were produced by the same talker or by two different talkers. The stimuli consisted of monosyllabic words that were presented in either matched-language (i.e., both English or both German) or mixed-language (i.e., English–German or German–English) pairs.

If indexical processing is language independent, then listeners should be able to discriminate talkers regardless of the language in which they are speaking. If indexical processing is language dependent, however, then language could affect performance in the voice discrimination task in at least two different ways. Listeners might, for instance, discriminate voices better when they are speaking in a familiar language. For native English listeners, voice discrimination would therefore be facilitated for English stimuli, yielding the best performance in English–English pairs, followed by the English–German and German–English pairs, and worst for the German–German pairs. Alternatively, language could affect voice discrimination performance in a different way if listeners attend to language-dependent indexical properties of speech regardless of the language they are listening to. In this case, performance in a discrimination task should be better in matched-language conditions (English–English and German–German) than in mixed-language conditions (English–German and German–English). Mixed-language conditions would force listeners to recalibrate their perceptual orientations between stimuli in order to attend to different sets of indexical properties in different languages. Not having to perform a similar recalibration between languages should facilitate discrimination accuracy in the matched-language conditions.

Not all of the listeners need to perform the voice discrimination task in the same way, of course. Some might process indexical information in a language-independent fashion while others might discriminate voices on the basis of language-dependent indexical cues. The results of experiment 1 provided evidence for both language-independent and language-dependent indexical processing depending on the language in which listeners were trained to identify voices. For this reason, both the English-trained and German-trained listeners from experiment 1 were brought back to participate in experiment 2 to determine if the perceptual proclivities

they had developed in learning to identify voices transferred to a voice discrimination task. A group of untrained English listeners also participated in experiment 2 in order to determine whether listeners without any experience with the particular bilingual talkers' voices would perform the discrimination task in a language-dependent or language-independent fashion. Comparing the discrimination performance of these listeners to that of the trained listeners also provided a means of determining how much voice identification training improved the ability of the listeners to discriminate between voices.

A. Methods

1. Stimulus materials

The stimuli for experiment 2 were produced by the same set of bilingual talkers that produced the stimuli for experiment 1.

2. Listeners

Three groups of listeners participated in experiment 2: English-trained, German-trained, and untrained listeners. The trained listener groups included 15 of the 20 listeners from each training group in experiment 1. These trained listeners were paid \$10 for their participation. Twenty-three additional listeners participated as untrained listeners. These listeners were students in undergraduate psychology courses at Indiana University who received partial course credit for their participation. Two untrained listeners were eliminated due to experimenter error and one was eliminated because of previous experience with German, resulting in 20 listeners in the untrained group. The remaining untrained listeners met the same qualifications as the trained participants: they were right handed, had no previous experience with German, had no history of a speech or hearing disorder, and were 18–25 years of age.

3. Procedure

Participants were tested in a quiet room. A customized software running on PCs presented the word pair stimuli to the listeners over Beyer Dynamic DT-100 headphones at a comfortable listening level. The participants were instructed to judge whether the two words in each pair were spoken by the same talker or by two different talkers. Participants registered their responses by pressing one of the two buttons on a custom-made button box; the right button registered "same" responses while the left button registered "different" responses. Participants were instructed to keep fingers on both buttons at all times during the experiment. They were also informed that the words they would be hearing might be spoken in an unfamiliar language.

Testing consisted of two blocks of 320 trials each. The stimuli in each block were evenly split between same-talker and different-talker trials. For the same-talker trials, word pairs were constructed by matching four pairs of different words produced by the same talker in four different language conditions: English–English, English–German, German–English, and German–German. For the different-talker trials, each talker was presented in combination with every other

FIG. 2. Box plot of sensitivity (A') values for all of the three listener groups in each language pair condition in experiment 2. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

same-gender talker twice—counterbalanced for order—in each of the four different language conditions. Cross-gender pairs were not presented to listeners as exceedingly few cross-gender confusions had been made in experiment 1 (approximately 1 out of every 300 identification trials). The two target words in the matched-language conditions were always different lexical items. For each listener, a different set of word pairs was selected from the database and trials were presented in a uniquely random order.

The words in each stimulus pair were separated by a 750 ms ISI. Listeners were instructed to respond as quickly as possible to each stimulus pair while still remaining accurate. Each block of trials began with a short sequence of four practice trials. After each of the trial in both the practice and testing sessions, the participants were informed whether their response was correct by a color-coded message (red for incorrect, blue for correct) presented on the computer screen. In testing, this message also informed the participants whether the pair had been a same-talkers trial or a different-talkers trial, along with the cumulative percentage of correct responses. Testing was self-paced, but participants generally completed each block of trials within 30 min. Participants were required to take a short break between blocks.

4. Data analysis

The same/different responses given by each listener were converted into nonparametric measures of sensitivity (A') and bias (B'') (Grier, 1971). Both of these measures are based on the proportion of “hits” and “false alarms” given by the listeners. Hits and false alarms were defined with respect to the same-talkers trials; a hit was a same response to a same-talkers trial, while a false alarm was a same response to a different-talkers trial. A' yields a measure of listener sensi-

tivity to the same-talkers/different-talkers distinction which ranges from 0.0 to 1.0, where a value of 1.0 indicates perfect discrimination and a value of 0.5 reflects chance performance on the discrimination task. B'' yields a measure of listener bias toward one response option or another. This measure ranges from -1.0 to 1.0 , where negative values indicate a tendency to give same responses, while positive values reflect a tendency to give different responses. A B'' value of 0 indicates lack of bias. Separate A' and B'' values were calculated for the responses given by each listener in each of the four different language pair conditions.

B. Results

One-sample t tests revealed that the sensitivity measures for all of the listener groups in all language pair conditions were significantly above 0.5, the level of chance performance of A' [for English-trained listeners, all $t(14) > 28.1$, $p < 0.001$; for German-trained listeners, all $t(14) > 36.6$, $p < 0.001$; for untrained listeners, all $t(19) > 24.1$, $p < 0.001$]. The average and range of the sensitivity values for all of the participants in each listener group are shown in Figure 2.

Repeated measure ANOVAs with language pair (English–English, German–German, English–German, German–English) as a within-subject factor and listener group (English-trained, German-trained, untrained) as a between-subject factor were conducted on both the sensitivity (A') and bias (B'') measures. The sensitivity ANOVA yielded main effects of listener group [$F(2,47) = 17.81$, $p < 0.001$] and language pair [$F(3,141) = 20.28$, $p < 0.001$], with no significant interaction between the two factors. *Post hoc* Tukey tests of the main effect of listener group revealed that both trained groups had significantly higher sensitivity than the untrained group ($p < 0.001$), but that there was no

TABLE II. Mean sensitivity (A') for the four listening conditions for all of the listener groups in experiment 2.

		Language pair				Mean
		FE	EG	GE	GG	
Listener group	English trained	0.891	0.864	0.843	0.867	0.866
	German trained	0.892	0.870	0.865	0.877	0.876
	Untrained	0.850	0.800	0.795	0.812	0.814
	Mean	0.878	0.845	0.834	0.852	0.852

significant difference in sensitivity between the two groups of trained listeners.

Post hoc paired sample t tests of the main effect of language pair revealed that listeners were better able to discriminate talkers in the English–English condition than in all other conditions (all $p < 0.001$). A significant difference between the German–German ($A' = 0.85$) and German–English ($A' = 0.83$) conditions was also found ($p = 0.007$), indicating that listeners could better discriminate talkers in the German–German condition. Table II provides the mean sensitivity values for each listener group in each of the language pair conditions.

The ANOVA on B'' yielded significant main effects of listener group [$F(2, 47) = 4.056$, $p = 0.024$] and language pair [$F(3, 141) = 46.32$, $p < 0.001$], as well as a significant interaction between the two factors [$F(6, 141) = 6.90$, $p < 0.001$]. *Post hoc* Tukey tests of the main effect of listener group revealed a significant difference between the biases of the German-trained listeners and the untrained listeners ($p = 0.021$). Untrained listeners were more biased to give same responses ($B'' = -0.123$) than the German-trained listeners ($B'' = -0.016$). Paired-sample t tests of the main effect of language pair revealed that all listeners were more likely to give same responses in the German–German condition than in the other three conditions (EE, $p = 0.009$; EG, $p < 0.001$; GE, $p < 0.001$). The listeners were also more likely to give same responses in the English–English condition than in the two mismatched-language conditions (both $p < 0.001$). Response bias did not differ significantly between the two mismatched-language conditions. Table III lists the mean response bias for each listener group in each of the language pair conditions.

The significant interaction between language pair and listener group on response bias is illustrated in Figure 3. *Post hoc* analyses revealed that this interaction was due to differences in response bias between the German-trained listeners and the other listener groups in both the English–German and German–German conditions. For these language pairs, the German-trained listeners were less likely to give same

responses ($p \leq 0.005$ in the English–German condition, and $p = 0.002$ in the German–German condition).

C. Discussion

The sensitivity values from the different language-pair conditions in experiment 2 indicate that the listeners performed the voice discrimination task by relying on both language-dependent and language-independent indexical information in the speech signal. For all of the three listener groups in all four language pair conditions, discrimination accuracy was significantly better than chance. Thus, listeners can accurately discriminate voices regardless of the language in which those voices were speaking. Voice discrimination was significantly better than chance even in the mismatched-language conditions (English–German and German–English). Since all listeners—including the untrained participants—exhibited this robust pattern, discrimination ability cannot solely depend on experience with either a particular language or a particular talker's voice.

The sensitivity values also indicate, however, that the language in which stimuli were presented did have some effect on the listeners' ability to discriminate talkers' voices. The listeners were better at discriminating voices in the matched-language conditions—especially the English–English pairs—than in the mixed language conditions. This pattern of language-dependent effects on sensitivity largely supports the hypothesis that listeners attend to different language-dependent indexical properties in different languages and must therefore recalibrate their perceptual orientations when listening to mixed language pairs. That the listeners showed better discrimination accuracy in the English–English condition than in the German–German condition is also consistent with the hypothesis that listeners can process language-dependent indexical information better in a familiar language.

The effects of language on sensitivity did not interact with the listeners' previous experience with the talkers; both of the English- and German-trained listeners performed sig-

TABLE III. Mean bias (B'') for the four listening conditions for all of the three listener groups in experiment 2.

		Language pair				Mean
		FE	EG	GE	GG	
Listener group	English trained	-0.093	-0.067	0.056	-0.291	-0.098
	German trained	-0.162	0.113	0.043	-0.056	-0.016
	Untrained	-0.144	-0.047	-0.024	-0.277	-0.123
	Mean	-0.133	0.000	0.025	-0.208	-0.079

FIG. 3. Box plots of bias (B'') values for all three listener groups in each language pair condition in experiment 2. Negative values reflect a bias toward same responses; positive values reflect a bias toward different responses. Means are indicated by a dark line in each box, and the length of each box represents 50% of the data. Whiskers extend to the largest and smallest values for each score; circles represent outliers.

nificantly better on the voice discrimination task than the untrained listeners. These trained listeners were therefore able to successfully transfer their knowledge of the talkers' voices across experimental tasks. Interestingly, there were no significant differences in sensitivity between the German- and English-trained listeners in any of the language pair conditions even though the generalization data from experiment 1 suggested that the German-trained listeners had processed indexical information in a more language-independent fashion than the English-trained listeners.

This absence of an interaction between training and language condition implies that the German-trained listeners' experience with hearing the bilingual talkers speaking in German in experiment 1 did not provide them with any additional advantage over the other two groups of listeners in processing indexical information in German. The analysis of the response bias revealed a different pattern of results between the listener groups, however. The listeners from all of the three groups were more biased to give same responses to matched-language pairs than to mixed-language pairs. In other words, the listeners were more biased to respond same when both stimulus words were spoken in the same language. This pattern indicates that listeners perceptually conflated linguistic and indexical information by interpreting words in the same language as coming from the same talker even when they were spoken by two different people. Although listeners were supposed to be performing a strictly indexical task, these bias measures indicate that they based their responses in part on the linguistic information in the signal regardless of the indexical content of the stimuli.

Within the matched-language conditions, the same response bias was stronger for the German–German pairs than for the English–English pairs. All of the listeners in this ex-

periment were thus more likely to conflate same-language information with same-voice responses when the talkers were speaking in German. The German-trained listeners, however, showed significantly reduced bias toward same responses in the German–German and English–German conditions than did the English-trained and untrained listeners. The tendency to perceptually conflate linguistic and indexical information was therefore less pronounced in the German-trained listeners. Training in English, on the other hand, did not change the listeners' response biases from their untrained counterparts. Both the English-trained and untrained listeners showed a strong same bias for the same language pairs and less same bias for the different language pairs. Thus, learning to identify voices in English improved the listeners' sensitivity to distinctions between voices but did not alter their perceptual biases; training in German, however, both improved the listeners' sensitivity and changed their response biases, such that these listeners showed a significantly reduced tendency to conflate linguistic and indexical information in perception.

It is interesting to note that the reduced same response bias by the German-trained listeners was limited to those language pairs which ended with German words. It is not entirely clear why this is the case. Had training in German simply enabled the listeners to better separate linguistic and indexical information in German words, a significant change in bias should have occurred for the German-trained listeners in the German–English language condition as well. Evidently, presenting the final word to the listeners in German triggered whatever language-independent processing abilities they had developed in training. Hearing the final word in English, on the other hand, caused them to revert to a more nativelike, language-dependent processing mode. The trans-

formation of these listeners' perceptual orientations through training in German was not, therefore, complete but rather continued to depend—to some extent—on the language of the stimuli that they heard.

IV. GENERAL DISCUSSION

One of the primary findings of these studies is that there is sufficient language-independent information in the speech signal to reliably identify and discriminate bilingual talkers' voices across two different languages. The presence of this information enabled the listeners to generalize their knowledge of talkers' voices across languages in experiment 1 and to accurately discriminate voices across languages in experiment 2. These findings support the results of previous research showing that the linguistic and indexical properties of speech are processed independently of one another. However, the results of this study also revealed that, in addition to the language-independent indexical properties of speech, listeners rely on language-dependent information to perform indexical tasks such as voice identification or discrimination. The listeners who were trained to identify talkers in English failed to generalize all of their knowledge of the talkers' voices from English to German in experiment 1; moreover, the discrimination performance of all of the groups of listeners was affected by the language in which stimuli were presented in experiment 2. These influences of language on indexical processing support the results of previous studies which have shown that listeners process the linguistic and indexical properties of speech in an integrated manner in perception. The combined results of the two experiments in this study thus validate both sides of the paradoxical findings of previous research—listeners apparently process the indexical properties of speech in both a language-dependent and a language-independent manner.

An exhaustive analysis of the acoustic properties that listeners used to identify and discriminate voices is beyond the scope of this paper, but the data do provide some clues as to what the most salient language-independent and language-dependent indexical cues were. One potential acoustic cue to talker identity that participants in the identification task consistently mentioned that they listened for was the “pitch” of each talker's voice. This was not a misguided strategy as most of the talkers used a characteristic *F0* pattern when recording the words for the bilingual talker database. Some of the talkers used a consistently high or low average *F0*, while others used a characteristic up-sloping or down-sloping intonation. One talker produced each item with a very short duration. These suprasegmental patterns were largely unique to each talker and were not determined by the phonological or semantic content of any of the words they produced. As such, they are not language-dependent properties of speech but rather reflect a pattern of articulatory choices made by each talker during the production of items in the recording session. For that reason, these talker-specific patterns carried over, to a large extent, across languages, making them a potential language-independent cue to talker identity which is not simply based on each talker's vocal tract physiology.

TABLE IV. Percentage of talkers correctly identified by the English listeners in the German generalization testing by German word type.

Word type	Examples	Total correct	Total heard	% correct
/x/	<i>Bach, doch</i>	91	176	51.7
/ts/	<i>Zeit, zahm</i>	55	108	50.9
Final /l/	<i>null, Ball</i>	131	260	50.4
English words	<i>Pein, nun</i>	198	401	49.4
English nonwords	<i>heiss, lahm</i>	385	791	48.7
Final /r/	<i>nur, Tier</i>	112	230	48.7
/pf/	<i>Pfeil, Kopf</i>	29	66	43.9
Initial /r/	<i>Rausch, Reim</i>	81	203	39.9
/o/, /e/	<i>Tod, Weg</i>	55	141	39.0
Front rounded vowels	<i>schoen, kuehl</i>	44	124	35.5

An examination of identification accuracy in the German generalization test of experiment 1 suggests that English-trained listeners identified talkers on the basis of language-specific segmental cues as well. Table IV presents the English-trained listeners' response accuracy in German generalization for the various phonetic word types in the German database; these word types are listed in descending order of talker identification accuracy. Word types include two categories for German words that are phonetically similar to English words—including both “words,” such as “mein” (“mine”) and “nun” (“noon”), and nonwords, such as “heiss” [hais] and “lahm” [lam]. There were also eight different word types for German words that included phonetic content not found in (General American) English: words with the velar/palatal fricative /x/, the “clear” /l/, front rounded vowels (/y/ and /ø/), initial /r/, final /r/, the affricates /pf/ and /ts/, and the mid- to high monophthongs /e/ and /o/.

It is difficult to draw clear conclusions about the data from this small sample set due to differences in the number of presentations for each word type, but a few general patterns emerge among the percentages. First, listeners had the most difficulty in identifying talkers from words that included vowels not found in English. Identification accuracy for words with front rounded vowels was only 35.5%, and identification accuracy for words with high to midmonophthongs was only 39%. The English-trained listeners were considerably less troubled by exotic consonant sounds, such as the velar fricative (51.7%) or the clear German /l/ (50.4%). Interestingly, the potential English word status of the German tokens did not seem to matter much to talker identification accuracy; possible English words such as “nun” and “mein” induced an accuracy level of 49.4%, while the talkers of possible English nonwords (heiss, lahm) were correctly identified 48.7% of the time. Numbers such as these indicate that, no matter how much English-trained listeners might have relied on language-independent cues such as patterns in *F0* to identify talkers, their talker identification responses were still sensitive to language-specific segmental information in the speech signal. In particular, the unfamiliar vowels of the German language caused problems for the listeners in generalization testing, suggesting that the English-trained listeners may have heavily relied on the acoustic information in language-specific vowel categories to identify

talkers during training. The interaction of language with indexical processing at this segmental level corroborates the hypothesis of [Remez et al., \(1997\)](#) that segmental phonetic content subserves both talker identification and word identification and is the locus for potential interactions between the two types of processing.

In postexperiment debriefing sessions, the English-trained listeners also indicated that they consciously attended to global indexical properties which would not be likely to transfer across languages. For instance, many listeners cited the relative accentedness of each talker's speech as a feature they listened to in trying to identify talkers during training. Another listener claimed that she could reliably identify one talker by the way that talker had "overexaggerated" the pronunciation of each word—a tendency, in other words, to "hyperarticulate" ([Lindblom, 1990](#)). Another listener reported that she could consistently identify one of the male speakers by the fact that he sounded "gay." Even if attending to such stylistic and sociophonetic markers of identity helped these listeners learn to identify the individual talkers' voices during training (there is no evidence for any such facilitation in the training scores), it is unlikely that any of these indexical properties would be available to listeners in a different language. 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 speaking German. A tendency to attend to language-dependent global and segmental properties of speech may therefore account for why the English-trained listeners' identification accuracy dropped when generalizing across languages in experiment 1.

In fact, a closer examination of the results of both studies suggests that the English-trained listeners—whether they consciously intended to or not—attended more closely to language-dependent indexical cues than the German-trained listeners in both tasks. The German-trained listeners were more successful in generalizing their knowledge of the talkers' voices across languages in experiment 1 and also showed a reduced bias toward perceptually conflating linguistic and indexical information in experiment 2. It is likely that the German-trained listeners learned to more heavily rely on language-independent information to perform indexical tasks simply because they were exposed to bilingual talkers speaking a language that they did not understand. Under these conditions, listeners could not be distracted by the linguistic content of the utterance and therefore learned to primarily attend to the language-independent indexical information in the signal. It is also possible that the German-trained listeners developed more abstract, language-independent representations of the talkers' voices simply because they were exposed to a new source of variability in speech (i.e., the German language). Effects of high stimulus variability in training that lead to better generalization performance in testing have been documented in a variety of earlier studies ([Bradlow et al., 1997](#); [Clopper, 2004](#); [Greenspan et al., 1988](#); [Iverson et al., 2005](#), [Logan et al., 1991](#)). While training in German did not necessarily involve more variability than training in English, it did present listeners with a different kind of acoustic-phonetic variability from their pre-

experiment experiences. Learning to process indexical information in this new linguistic form may therefore have better guided the listeners' perceptual systems in attending to the information in the signal which was most relevant to the indexical processing task rather than to any distracting linguistic properties of the signal.

The poorer performance of the English-trained listeners on the cross-linguistic generalization task in experiment 1 suggests that processing indexical information in a language-dependent fashion diverts attention and makes the generalization of indexical knowledge across languages more difficult. If this is the case, then why did the listeners do it? One answer is that language-dependent indexical cues may provide listeners with some indexical processing advantages, as it did for all of the groups of listeners in the English-English condition of experiment 2. Another answer may be that the listeners simply cannot help themselves—the language-dependent interpretation of indexical information in speech may be an automatic perceptual process, with negative processing consequences that are only made apparent when listeners are introduced to new types of linguistic variability in the signal. Therefore, listeners may automatically attend to language-dependent indexical cues in speech when they are listening to a language that they understand. Interestingly, the results of this study show that this process may take place even though reliance on language-dependent cues is not necessary for either talker identification or discrimination. There is sufficient language-independent information in the signal to support both talker identification and discrimination across different languages.

The different behaviors of the English-trained and German-trained listeners in these experiments have several implications for exemplar theories of speech perception, which originally drew inspiration from findings that listeners store in memory all tokens of experienced speech, without segregating indexical and linguistic properties in long-term representations. The results of this study suggest that not all exemplars are created equal. The English-trained listeners exhibited language-dependent processing in both experiments, indicating that they had developed integrated indexical and linguistic representations in memory. The German-trained listeners, on the other hand, seemed to develop largely language-independent representations of talkers' voices. If so, these listeners would not be expected to show an interaction between linguistic and indexical properties when processing English words spoken by familiar bilingual talkers. An advantage for identifying English words spoken by familiar talkers has been shown for English-trained listeners ([Nygaard et al., 1994](#)), but preliminary evidence from our laboratory indicates that German-trained listeners do not exhibit such a familiar talker advantage across languages ([Levi et al., 2007a](#)). The integrated representations in memory that are generally assumed by exemplar theory therefore only seem to emerge when listeners know how to interpret both the linguistic and indexical information in the signal.

There are mechanisms in place in some exemplar-based models of speech perception which can account for this pattern of findings. [Pierrehumbert \(2002\)](#), for instance, suggests that exemplar-based storage requires knowing how to assign

the different kinds of variability found in speech to specific category labels. Unassigned variation is simply discarded from memory. The German-trained listeners in this study who lacked category labels for German words and German-specific phonemic categories may thus have simply excluded from memory the variation in the signal that would be relevant to German language processing. Alternatively, the two groups of listeners may have differed in the amount of attention they devoted to the linguistic properties of the signal in perception. Computationally based models of exemplar categorization, such as Kruschke's *ALCOVE* (1992) or Johnson's *XMOD* (1997), typically incorporate attention weights into a front-end processing system to support exemplar storage in memory. In speech perception, particular linguistic properties may be enhanced or compressed in the stored exemplar representation according to the amount of attention listeners pay to them (e.g., Escudero, 2005; Iverson *et al.*, 2005; Kruschke, 1992; Nosofsky, 1986). In the case of the current study, the English-trained listeners may simply have paid more attention to the language-specific properties in the signal, while the German-trained listeners focused more on the language-independent properties of the talkers' voices. The different attention weights thus led to different representations in the stored exemplars themselves.

It should be noted here that the evidence for exemplar theories of speech perception largely (if not exclusively) come from studies in which listeners were presented with familiar language stimuli. It is not necessarily true that listeners process unfamiliar languages in the same way. The results of this study suggest that more consideration ought to be given to how exemplars are encoded in perception and what, exactly, is stored in memory. The findings of these experiments suggest that investigating the potential interactions of indexical and linguistic information across languages—rather than just in English—offers a promising avenue for future research on this increasingly influential theory.

V. CONCLUSION

The combined results of the two experiments reported here suggest that the extent to which indexical processing relies on language-independent information depends on the listeners' knowledge of the language presented in the speech signal. Listeners process indexical information in a language-dependent fashion when they hear a language that they know; otherwise, they perform indexical tasks by more heavily relying on language-independent information in the signal. This general perceptual pattern suggests that listeners shift their focus of attention, in fundamental ways, to adapt to the linguistic structure of the speech signals they receive and the specific processing demands of the perceptual task.

The finding that the human perception of speech can rapidly adapt to changing linguistic conditions in this way may reconcile the apparently conflicting evidence for both the language-dependent and language-independent views of speech perception that were presented in the introduction. Indexical processing operates in a language-dependent manner whenever listeners understand the language that is being

spoken. In this case, they can use integrated linguistic and indexical information in the speech signal to identify or discriminate voices. For this reason, listeners identify voices better when they are speaking in a familiar language, but they also have difficulty generalizing their knowledge of voices from a familiar to an unfamiliar language. On the other hand, when linguistic information is removed from the speech signal—as in filtered speech or phonagnosia or by presenting stimuli in an unfamiliar language—the perceptual system accordingly adjusts by relying on the language-independent information in the signal to perform indexical perception tasks. The perception of the indexical properties of speech may thus be either language-dependent or language-independent depending on the context in which listeners operate. 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.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health to Indiana University (NIH-NIDCD T32 Training Grant No. DC-00012 and NIH-NIDCD Research Grant No. R01 DC-00111). The authors would like to thank Christina Fonte, Jen Karpicke, and Melissa Troyer for their help in running the subjects and editing stimuli. The authors would also like to thank two anonymous reviewers for their comments on an earlier draft of this paper.

¹Talker-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 in a training paradigm are presented to them more than once (Winters *et al.*, 2005).

²The p -value ($p=0.049$) approaches significance with a Bonferroni correction to $p=0.0125$.

³This finding includes a Bonferroni correction to $p=0.00625$ for the eight comparisons made on these data. For sessions 5, 7, and 8, $p>0.10$. For session 6, $p=0.038$ for the German generalization and $p>0.10$ for the English generalization.

Abercrombie, D. (1967). *Elements of General Phonetics* (Edinburgh University, Edinburgh).

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX Lexical Database, Release 2 (CD-ROM), Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bradlow, A. R., Pisoni, D. P., Akahane-Yamada, R., and Tohkura, Y. (1997). "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production," *J. Acoust. Soc. Am.* **101**, 2299–2310.

Bricker, P. D., and Pruzansky, S. (1966). "Effects of stimulus content and duration on talker identification," *J. Acoust. Soc. Am.* **40**, 1441–1449.

Clarke, F. R., Becker, R. W., and Nixon, J. C. (1966). "Characteristics that determine speaker recognition," Report No. ESD-TR-66-638, Electronic Systems Division, Air Force Systems Command, pp. 1–65, Hanscom Field, MA.

Clopper, C. G. (2004). "Linguistic experience and the perceptual classification of dialect variation," Ph.D. thesis, Indiana University, Bloomington, IN.

Compton, A. J. (1963). "Effects of filtering and vocal duration upon the identification of speakers, aurally," *J. Acoust. Soc. Am.* **53**, 1741–1743.

Escudero, P. (2005). "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," Ph.D. thesis, Utrecht University, Utrecht, the Netherlands.

Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (2003). "Consolidation

- during sleep of perceptual learning of spoken language," *Nature (London)* **425**, 614–616.
- Glisky, E. L., Polster, M. R., and Routhieaux, B. C. (1995). "Double dissociation between item and source memory," *Neuropsychology* **9**, 229–235.
- Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (1991). "The role of language familiarity in voice identification," *Mem. Cognit.* **19**, 448–458.
- Goldinger, S. D. (1996). "Words and Voices: Episodic traces in spoken word identification and recognition memory," *J. Exp. Psychol. Learn. Mem. Cogn.* **22**, 1166–1183.
- Goldinger, S. D. (1998). "Echoes of echoes? An episodic theory of lexical access," *Psychol. Rev.* **105**, 251–279.
- Goldinger, S. D., Pisoni, D. B., and Logan, J. S. (1991). "On the locus of talker variability effects in recall of spoken word lists," *J. Exp. Psychol. Learn. Mem. Cogn.* **17**, 152–162.
- Greenspan, S. L., Nusbaum, H. C., and Pisoni, D. B. (1988). "Perceptual learning of synthetic speech produced by rule," *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 421–433.
- Grier, J. B. (1971). "Nonparametric indexes for sensitivity and bias: Computing formulas," *Psychol. Bull.* **75**, 424–429.
- Grosjean, F. (1980). "Spoken word recognition processes and the gating paradigm," *Percept. Psychophys.* **28**, 267–283.
- Iverson, P., Hazan, V., and Bannister, K. (2005). "Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults," *J. Acoust. Soc. Am.* **118**, 3267–3278.
- Johnson, K. (2005). "Speaker normalization in speech perception," *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. Remez (Blackwell, Oxford), pp. 363–389.
- Johnson, K. A. (1997). "Speech perception without speaker normalization," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. Mullennix (Academic, San Diego), pp. 145–165.
- Köster, O., and Schiller, N. O. (1997). "Different influences of the native language of a listener on speaker recognition," *Forensic Linguistics* **4**, 18–28.
- Kruschke, J. (1992). "ALCOVE: An exemplar-based connectionist model of category learning," *Psychol. Rev.* **99**, 22–44.
- Landis, T., Buttet, J., Assal, G., and Graves, R. (1982). "Dissociation of ear preference in monaural word and voice recognition," *Neuropsychology* **20**, 501–504.
- Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., and Bourne, V. T. (1976). "Speaker sex identification from voiced, whispered, and filtered isolated vowels," *J. Acoust. Soc. Am.* **59**, 675–678.
- Levi, S. V., Winters, S. J., and Pisoni, D. B. (2007a). "A cross-language familiar talker advantage?," *Research on Speech Perception Progress Report No. 28*, Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 369–383.
- Levi, S. V., Winters, S. J., and Pisoni, D. B. (2007b). "Speaker-independent factors affecting the degree of perceived foreign accent in a second language," *J. Acoust. Soc. Am.* **121**, 2327–2338.
- Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory," *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), pp. 403–439.
- Logan, J. D., Lively, S. E., and Pisoni, D. B. (1991). "Training Japanese listeners to identify English /r/ and /l/: A first report," *J. Acoust. Soc. Am.* **89**, 874–886.
- Mullennix, J. W., and Pisoni, D. B. (1990). "Stimulus variability and processing dependencies in speech perception," *Percept. Psychophys.* **47**, 379–390.
- Nagao, K. (2006). "Cross-language study of age perception," Ph.D. thesis, Indiana University, Bloomington, IN.
- Nosofsky, R. M. (1986). "Attention, similarity, and the identification-categorization relationship," *J. Exp. Psychol. Gen.* **115**, 39–57.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophys.* **60**, 355–376.
- Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). "Speech perception as a talker-contingent process," *Psychol. Sci.* **5**, 42–46.
- Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993). "Episodic encoding of voice attributes and recognition memory for spoken words," *J. Exp. Psychol. Learn. Mem. Cogn.* **19**, 309–328.
- Pierrehumbert, J. (2001). "Exemplar dynamics: Word frequency, lenition, and contrast," in *Frequency effects and the Emergence of Lexical Structure*, edited by J. Bybee and P. Hopper (Benjamins, Amsterdam), pp. 137–157.
- Pierrehumbert, J. (2002). "Word-specific phonetics," in *Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), pp. 101–140.
- Pollack, I., Pickett, J. M., and Sumbly, W. H. (1954). "On the identification of speakers by voice," *J. Acoust. Soc. Am.* **26**, 403–406.
- Remez, R. E., Fellowes, J. M., and Rubin, P. E. (1997). "Talker identification based on phonetic information," *J. Exp. Psychol. Hum. Percept. Perform.* **23**, 651–666.
- Schacter, D. L., and Church, B. A. (1992). "Auditory priming: Implicit and explicit memory for words and voices," *J. Exp. Psychol. Learn. Mem. Cogn.* **18**, 915–930.
- Schiller, N. O., and Köster, O. (1996). "Evaluation of a foreign speaker in forensic phonetic: A report," *Forensic Linguistics* **3**, 176–185.
- Schiller, N. O., Köster, O., and 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," *Brain Res. Cognit. Brain Res.* **18**, 162–171.
- Sullivan, K. P. H., and 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," *Appl. Cognit. Psychol.* **1**, 121–131.
- Todaka, Y. (1993). "Japanese students' English intonation," *Bulletin of Miyazaki Municipal University* **1**, 23–47.
- Van Lancker, D. R., Cummings, J. L., Kreiman, J., and 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," *Report No. ESD-TDR-65-153*, Electronic Systems Division, Air Force Systems Command, Hanscom Field, MA.
- Winters, S. J., Levi, S. V., and Pisoni, D. B. (2005). "When and why feedback matters in the perceptual learning of the visual properties of speech," *Research on Speech Perception Progress Report No. 27*, Speech Research Laboratory, Indiana University, Bloomington, IN, pp. 107–132.