Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-019-01778-5

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Hierarchical contributions of linguistic knowledge to talker identification: Phonological versus lexical familiarity

Deirdre E. McLaughlin[1] · Yaminah D. Carter[1] · Cecilia C. Cheng[1] · Tyler K. Perrachione[1]

Abstract
Listeners identify talkers more accurately when listening to their native language compared to an unfamiliar, foreign language. This *language-familiarity effect* in talker identification has been shown to arise from familiarity with both the sound patterns (phonetics and phonology) and the linguistic content (words) of one's native language. However, it has been unknown whether these two sources of information contribute independently to talker identification abilities, particularly whether hearing familiar words can facilitate talker identification in the absence of familiar phonetics. To isolate the contribution of lexical familiarity, we conducted three experiments that tested listeners' ability to identify talkers saying familiar words, but with unfamiliar phonetics. In two experiments, listeners identified talkers from recordings of their native language (English), an unfamiliar foreign language (Mandarin Chinese), or "hybrid" speech stimuli (sentences spoken in Mandarin, but which can be convincingly coerced to sound like English when presented with subtitles that prime plausible English-language lexical interpretations based on the Mandarin phonetics). In a third experiment, we explored natural variation in lexical-phonetic congruence as listeners identified talkers with varying degrees of a Mandarin accent. Priming listeners to hear English speech did not improve their ability to identify talkers speaking Mandarin, even after additional training, and talker identification accuracy decreased as talkers' phonetics became increasingly dissimilar to American English. Together, these experiments indicate that unfamiliar sound patterns preclude talker identification benefits otherwise afforded by familiar words. These results suggest that linguistic representations contribute hierarchically to talker identification; the facilitatory effect of familiar words requires the availability of familiar phonological forms.

Keywords Speech perception · Talker identification · Phonology · Priming · Accent · Language-familiarity effect

Introduction

Talker identification – the process of identifying a speaker by the sound of their voice – is an important social and perceptual skill. Research has consistently demonstrated that the ability to identify talkers is functionally integrated with the processes involved in perceiving speech. A prominent phenomenon demonstrating this integration is the *language-familiarity effect* in talker identification, in which listeners are more accurate at identifying talkers by the sound of their voice when listening to speech in their native language compared to unfamiliar or foreign languages (Goggin et al., 1991; Perrachione & Wong, 2007; Thompson, 1987). This effect of language on processing talker identity underscores a bi-directional relationship between linguistic and social-perceptual faculties (Kuhl, 2011): Listeners are able to both resolve talker variability in order to arrive at an underlying linguistic message (e.g., Choi, Hu, & Perrachione, 2018; Mullennix & Pisoni, 1990) and employ an underlying linguistic representation in order to more accurately identify a speaker by the sound of their voice (e.g., Perrachione, Del Tufo, & Gabrieli, 2011).

Although the relationship between language familiarity and talker identification ability is reliably observed in a large body of scientific work (reviewed in Perrachione, 2018), there

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13414-019-01778-5) contains supplementary material, which is available to authorized users.

✉ Tyler K. Perrachione
tkp@bu.edu

[1] Department of Speech, Language, and Hearing Sciences, College of Health and Rehabilitation Sciences: Sargent College, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA

Published online: 19 June 2019

remains no agreed-upon cognitive model to explain either *what* information is integrated between these two faculties or *how* such integration occurs. Some authors have asserted a role for higher-level linguistic processing in talker identification, in which listeners gain access to talker identity-relevant information by processing and representing speech at the level of familiar linguistic units such as words (e.g., McLaughlin et al., 2015; Perrachione, Del Tufo, & Gabrieli, 2011). Other authors have described how the language-familiarity effect can arise from acoustic-phonetic processing, in which listeners gain access to talker identity-related information by processing speech with respect to the familiar phonetic patterns of their native language (Fleming et al., 2014; Zarate et al., 2015). Although both sources of information – acoustic-phonetic and lexical – have been found to simultaneously facilitate native-language talker identification (Perrachione et al. 2015), it is currently unknown whether these sources of information contribute independently to this ability, or whether there is a bi-directional or hierarchical dependence between these representations. Determining when and how various levels of linguistic knowledge affect talker identification is necessary to better understand the integration between linguistic, perceptual, and mnemonic processes in the human mind.

The role of familiar sounds and sound patterns in talker identification

Several lines of evidence support the idea that speech with familiar phonetics and phonological structure facilitates listeners' perception of talker identity, even when that speech lacks familiar words. Listeners identify talkers more accurately from meaningless pseudo-words that follow the sound structure of their native language than they do from foreign-language speech (Perrachione et al., 2015; Xie & Myers, 2015b; Zarate et al., 2015), suggesting familiar sound structure gives listeners access to additional talker-specific information even in the absence of comprehensibility. The benefits of sound-structure familiarity may not even depend on linguistic structure derived from word knowledge: When learning to identify talkers speaking in French, self-reported monolingual English listeners from Canada outperformed monolingual English listeners from the USA, suggesting that incidental, passive exposure to the sound structure of an unfamiliar language may also facilitate talker identification, even in the putative absence of any familiar word forms (Orena, Theodore, & Polka, 2015). Likewise, for infants as young as 7 months, familiarity with the sound structure of their native language, even though they recognize few if any words, is sufficient to elicit a form of the language-familiarity effect (Fecher & Johnson, 2018b; Johnson et al., 2011).

The idea that accurate talker identification is driven in part by phonological familiarity is supported by some reports that show a larger language-familiarity effect between languages that are more dissimilar phonologically (Zarate et al., 2015), although such effects of phonological dissimilarity are not commonly reported (Johnson et al., 2011; Köster & Schiller, 1997; Xie & Myers, 2015a). Finally, when listening to time-reversed speech (which putatively retains certain acoustic-phonetic features while rendering speech incomprehensible) listeners rate talkers of their native language as more dissimilar sounding than talkers of a foreign language (Fleming et al., 2014; cf. Furbeck et al., 2018), suggesting that listeners are more sensitive to inter-talker differences in the presence of familiar, language-specific acoustic structure. Collectively, there is an abundance of evidence that listening to speech with familiar acoustic-phonetic properties contributes to more accurate processing of talker-identity related information.

The role of familiar words and higher-level linguistic units in talker identification

There is also evidence demonstrating that, beyond familiarity with sound structure, talker identification is facilitated by higher-level linguistic processing, particularly representations at the level of words. Several studies have shown that talker identification abilities improve as a function of the amount of linguistic information available from talkers. Listeners identify talkers more accurately as their speech increases in complexity from vowels to words to sentences (Bricker & Pruzansky, 1966; Goggin et al., 1991; Pollack, Pickett, & Sumby, 1954) – an effect that appears to hold for foreign-accented speech as well (Goldstein, Knight, Bailis, & Conover, 1981). When listening to two-word sequences, listeners detect a change in talker across words more accurately when the two words are unrelated than when the words form a meaningful sequence, demonstrating integrated processing of lexical-semantic and phonetic-indexical information (Narayan, Mak, & Bialystok, 2017). Talker identification is also improved as the quantity of known, as opposed to novel, words increases: Listeners perform better when identifying talkers from speech comprised of real words compared to nonsense speech matched in native-language phonological structure (Goggin et al., 1991; Perrachione et al., 2015; Xie & Myers, 2015b). Listeners also learn to identify talkers more accurately in their native language, but not a foreign language, when the lexical content of the speech is repeated, revealing that consistent (but unknown) speech content confers no talker identification benefit in a foreign language, whereas listeners' ability to identify native-language voices improves with their ability to remember and compare the content of their speech (McLaughlin et al., 2015).

Different task demands can also highlight the comparative importance of different levels of linguistic representation during voice and talker perception. Whereas perceptual dissimilarity ratings of time-reversed speech appear to be affected by

correspondence in the language spoken by talkers and listeners (Fleming et al., 2014), listeners do not appear to gain an advantage in the *identification* of talkers in their native language when recordings have been time-reversed (Perrachione et al., 2015). Similarly, whereas familiarity with the language spoken by talkers imparts a large and reliable advantage in talker identification (Perrachione, 2018), linguistic familiarity does not appear to give listeners as much of an advantage in *discriminating* whether two speech samples come from the same or different talkers (Fecher & Johnson, 2018a; Wester, 2012; Winters, Levi, & Pisoni, 2008). That more complex tasks, such as talker identification, increasingly draw upon higher-level representations compared to simpler tasks, such as talker discrimination, raises the possibility that there are additive contributions of various levels of linguistic knowledge in representing talker-specific information. What remains unknown is whether these levels of representation can contribute independently, or whether there is a hierarchical dependence between lower and higher levels of representation in encoding talker identity-related information.

The present study: Do familiar words always benefit talker identification?

Across three experiments, we explored whether talker identification abilities benefitted from processing familiar lexical information independently from familiar acoustic-phonetic information. Specifically, we examined whether being able to parse a speech stream into familiar words, particularly when the sound structure was unfamiliar, would nonetheless facilitate talker identification accuracy. In the first two experiments, listeners heard sentences spoken in Mandarin that could be convincingly coerced to sound like English when presented with subtitles that primed lexical expectations during speech processing. These sentences were carefully designed to create semantically and syntactically plausible sentences in both languages, with the presence of subtitles priming plausible English glosses of the Mandarin speech.

The coercion of speech produced in one language to sound convincingly like speech from another language has been widely demonstrated in the pop-culture phenomenon of "mondegreens," in which speech (frequently song lyrics) in a foreign language is heard as native-language speech in the presence of simultaneous native-language subtitles (Liberman, 2007). Speech perception research in the laboratory has likewise demonstrated numerous circumstances where top-down expectations about words influence listeners' speech processing. The classical example of biasing perception based on lexical expectations comes from the Ganong effect in categorical perception, in which listeners are biased to disregard competing acoustic information in favor of perceiving real words (Ganong, 1980). Biasing perception of speech based on listeners' expectations also extends to richer phonetic contexts such as sentences. Perception of vocoded sentences, where detailed spectral information is removed from the speech signal, is more accurate when listeners are primed to expect key content words from the sentence (Davis et al., 2005). Using subtitles to prime lexical expectations also helps listeners perceive the words in vocoded speech more accurately (Sohoglu & Davis, 2016). In perhaps the most compelling example of the power of expectations to bias perception in favor of real words, listeners report actually "hearing" target words in speech when they have been primed to expect those words, even when all distinguishing spectral and temporal information from the acoustic signal has been completely effaced (which renders speech otherwise totally incomprehensible; Holdgraf et al., 2016).

From the Ganong effect to the identification of vocoded speech, the power of top-down expectations to alter the correspondence between sensory input and linguistic representations is well established in speech processing. But can these top-down linguistic biases also affect the correspondence between sensory inputs and talker representations? Voice processing may take advantage of a perceptual space wherein talkers' voices are encoded as deviations from a prototype voice (Latinus & Belin, 2011), the specification of which likely depends on language-specific representations of voices (Goggin et al., 1991) constructed from language-specific acoustic, phonetic, phonological, and lexical features (e.g., Fleming et al., 2014). However, the phonetic-phonological correspondences differ across languages (e.g., Lisker & Abramson, 1964), and thus the informative variability in talker-specific phonetic idiosyncrasies may be more opaque to listeners when they are identifying foreign-language voices. Higher-level linguistic structure, such as words, guides both the perception and interpretation of ambiguous phonetic information (Getz & Toscano, 2019; Samuel, 1997, 2001) and can facilitate phonetic processing even in an unfamiliar language (Samuel & Frost, 2015). Correspondingly, by providing listeners with higher-level linguistic representations through which they can interpret the ambiguous phonetics of foreign language speech, known lexical content may give listeners a scaffold upon which they can extract more information about talker-specific phonetic variation and thus facilitate foreign-language talker identification.

In the present study, we first tested the hypothesis that priming listeners to parse a foreign-language speech stream comprised of unfamiliar sounds into real words via native-language subtitles would improve talker identification accuracy compared to a condition in which no primes were presented. If this manipulation improved talker identification from foreign-language speech, it would favor a model of talker identification in which facilitatory representations of voices are made available via lexical processing in parallel with talker-specific information provided by familiar sound structure. However, if allowing listeners to parse a speech stream

comprised of unfamiliar sounds into one made up of familiar words has no effect on talker identification, it would suggest that the talker identification benefits conferred by processing the lexical content of speech (e.g., Goggin et al., 1991; McLaughlin et al., 2015; Perrachione et al., 2015) are only available when the acoustic-phonetic features of speech are also familiar. This latter result would, instead, favor a model of talker identification in which the facilitatory contribution of familiar words has a hierarchical dependence on the availability of familiar sound structure.

In two versions of this experiment involving different amounts of training, we found that, contrary to our expectations, lexical priming does not appear to improve talker identification in the absence of familiar phonological information. The laboratory manipulation of coercing foreign-language speech with an unfamiliar phonology to sound like listeners' native language is somewhat analogous to the common, real-world situation of listening to speech with a heavy foreign accent. Thus, we ran a third, follow-up experiment in which we investigated a related hypothesis: that the degree of phonetic dissimilarity (operationalized here as the degree of perceived foreign accent) negatively affects talker identification abilities for speech produced in listeners' native language. In this experiment, we observed a graded effect of unfamiliar phonetics on English-speaking listeners' talker identification abilities, with most accurate talker identification for native English-accented talkers, followed by Mandarin-English bilinguals with a slight Mandarin accent (low-accentedness), Mandarin-English bilinguals with a stronger Mandarin accent (high-accentedness), and with Mandarin-speaking talkers identified least accurately. Taken together, the results from these three experiments strongly suggest that familiarity with the sound structure of speech has precedence over processing higher-level linguistic structure when conferring a benefit in talker identification, and thus that linguistic information contributes to talker identification in a hierarchical fashion, with higher levels of representation conferring a benefit only when lower levels are also familiar.

Experiment 1: Priming lexical representations during foreign-language talker identification

In this experiment, we investigated whether allowing listeners to parse a speech stream composed of unfamiliar sounds into familiar words via lexical priming with subtitles could confer a benefit in learning to identify talkers compared to listening to speech from the same foreign language without lexical priming. In a within-subjects, 2 × 2 factorial design, native English-speaking listeners learned to identify talkers speaking in either English or Mandarin, with or without accompanying subtitles to prime listeners to hear English words from the speech. Listeners completed each of these four talker identification conditions (English/Mandarin-speaking talkers presented with/without subtitles) separately in a counterbalanced order.

Methods

Participants

Native speakers of American-English completed this study (N = 32, 26 female, six male; age 18–35 years, M = 21.8). Inclusion criteria required participants to have a self-reported history free from speech, language, or hearing problems and no prior experience with Mandarin. This study was approved and overseen by the Institutional Review Board at Boston University. Participants provided written informed consent and were paid for their participation.

The sample size was determined by the number of permutations of experimental conditions necessary to counterbalance the stimuli, and is larger than most of the prior studies of the role of language in talker identification (Perrachione, 2018). Previous research found that manipulations involving lexical content in talker identification have effect sizes on the order of Cohen's d = 0.5-1.2 (McLaughlin et al., 2015; Perrachione et al., 2015). Correspondingly, with N = 32 we have 87% to 100% power to detect effect sizes in the published range, and 80% power to detect effect sizes of $d \geq 0.45$.

Stimuli

Twenty "English-Mandarin hybrid sentences" were designed for this experiment (Table 1 and Appendix). Each hybrid sentence was syntactically correct and semantically plausible in both languages, but the English and Mandarin forms of the

Table 1 Example Mandarin-English hybrid sentences used in Experiments 1 and 2

Mandarin	English gloss
陪你晚到了 /pʰeɪ ni wan tau lə/ péi nǐ wǎn dào le *With you, I was late.*	"Pay me one dollar." /pʰeɪ mi wʌn dɑlə/
喂狗吃烤荔枝 /wei kou tʰ kʰau li ʦi/ wèi gǒu chī kǎo lì zhī *Feed the dog grilled lychees.*	"We go to college." /wi goʊ tʰu kʰɑlədʒ/
妈妈喜欢帽子 /ma ma ɕi xuɑn mau ʦi/ mā mā xǐ huān mào zi *Mother likes the hat.*	"Mama sees one mouse." /mɑmə siz wən maʊs/

Mandarin versions of the hybrid sentences are written in simplified characters and accompanied by their phonetic transcription according to the International Phonetic Alphabet (IPA), their phonetic transcription in *pinyin*, and their literal translation in italics. The English glosses are shown in quotes, with the corresponding phonetic transcription in IPA

sentence were not translations between the two languages. Instead, the sentences – originally constructed in Mandarin – were designed to have an intended English "gloss" that could convincingly be heard from the phonetics of natural Mandarin speech. The Mandarin sentence and its English gloss were designed based on correspondences between the phonotactic properties of English, Mandarin, Mandarin-accented English, and the patterns of (mis)perception of Mandarin phonemes by English speakers (e.g., Tsao, Liu, & Kuhl, 2006). For example, in the hybrid sentence "陪你晚到了" (/pʰeɪ ni wɛn tau lə/), a listener expecting to hear Mandarin-accented English can convincingly hear, "Pay me one dollar" (/pʰeɪ mi wʌn dɑlə/), a mapping to English words that capitalizes on, among other features, reliable perception of the Mandarin voiceless but unaspirated [t] by English listeners as an English /d/ and the typical reduction in r-coloring of rhotic vowels in Mandarin-accented English. Hybrid stimuli were extensively piloted prior to use in this talker identification study to ensure they could elicit the intended English speech percept, particularly when presented with concomitant subtitles. We also confirmed that orienting listeners' perceptual expectations towards an English interpretation of the Mandarin speech was effective at eliciting the intended English glosses during the actual talker identification experiment through a supplemental sentence transcription task, undertaken by a subset of participants after completing the Mandarin conditions of the talker identification task. This stimulus validation is described in detail below. (Example audio recordings of the English-Mandarin hybrid stimuli used in Experiment 1 are available as Supplementary Materials.)

The English-Mandarin hybrid sentences were recorded (in Mandarin) by ten female native speakers of Mandarin (age 19–27 years, M = 23 years). Corresponding recordings (in English) of the hybrid sentences' intended English glosses were made by ten female native speakers of American English (age 19–29 years, M = 22.3 years). Both groups of talkers were without distinctive regional accents. Recordings were made in quiet in a sound-attenuated booth using a Shure MX153 earset microphone, a Behringer Ultragain Pro MIC2200 2-channel tube microphone preamplifier, and Roland Quad Capture USB audio interface with a sampling rate of 44.1 kHz and 16-bit digitization in Praat RMS amplitude. Each sentence was RMS-amplitude normalized to 65 dB SPL using Praat v5.3.63.

In the talker identification experiment, listeners learned to identify two sets of talkers in each language, once with subtitles accompanying their recordings, once with no subtitles. Because some voices are inherently more distinctive than others, we arranged our talkers in each language into two, five-voice sets that would be equally identifiable on average. Additional pilot listeners learned to identify various groupings of these voices, allowing us to calibrate listeners' within-language accuracy to be equal between the two sets of talkers. This piloting ensured that, absent of the lexical priming manipulation in the actual experiment, listeners' mean accuracy would not differ between repetitions of the talker identification task with different speakers of each language. Furthermore, the two sets of talkers in each language were also counterbalanced so they appeared equally often with or without accompanying subtitles.

Procedure

In a within-subjects, 2 × 2 factorial design experiment, participants learned to identify talkers across manipulations of the language being spoken (English or Mandarin) and the presence of top-down lexical priming (with or without subtitles), resulting in four conditions: (1) English with subtitles, (2) English without subtitles, (3) Mandarin with subtitles, and (4) Mandarin without subtitles. In order to preserve the illusion that the Mandarin with subtitles condition was actually English, before this condition participants were told that they were hearing English speech with a heavy Mandarin accent, and that subtitles were being provided to help them recognize the speech. Prior to the Mandarin-without-subtitles condition, participants were told they would be hearing speech in a foreign language they would not be able to understand. In all conditions, participants were also told that their ability to understand the speech was not important, that we were interested in their ability to learn to identify the talkers.

Participants completed all conditions of the experiment in a single session, and the order of conditions was counterbalanced across participants. Participants learned a unique group of five voices in each condition, and the speech content (i.e., which hybrid sentences were presented) was unique in each condition. The sentences used in each condition were permuted across participants, and the talkers used in the subtitle versus no-subtitle conditions were also permuted (within language) across participants, to control for voice and item effects on the experimental manipulations.

Talker identification training and testing

The procedure for talker identification training and testing was the same in each condition (Fig. 1A), excepting the manipulations of the language being spoken and the presence of subtitles. In each condition, listeners learned to associate five talkers with five unique, numbered avatars. First, listeners were familiarized with, and practiced identifying, the five talkers in a series of interleaved passive listening and active identification blocks. Following familiarization, listeners were tested on their ability to correctly identify the talkers.

In the training phase of each condition, participants learned to identify the talkers by the sound of their voice across five interleaved blocks of passive familiarization and active identification practice. This procedure has been used extensively

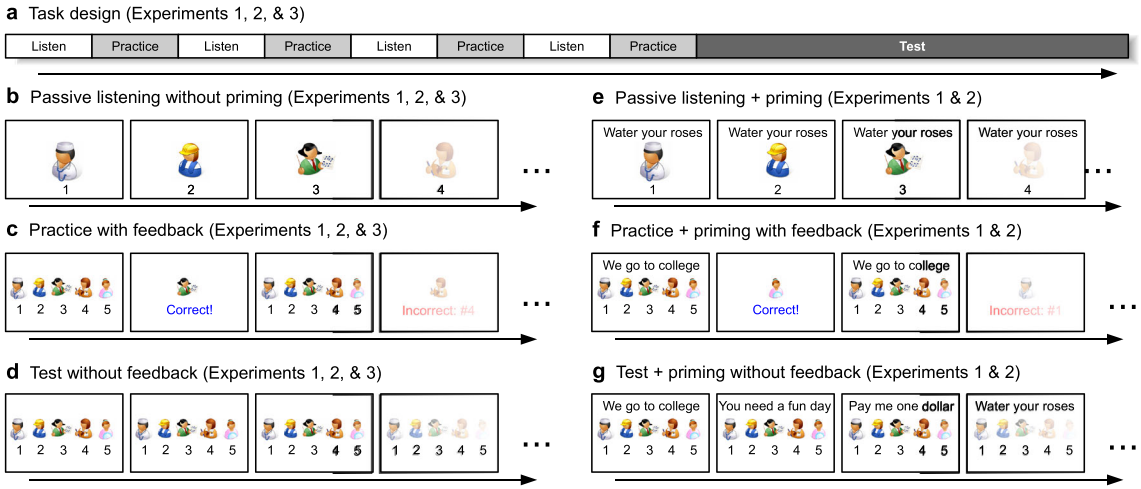

Fig. 1 Talker identification training and testing paradigm. (**A**) Across Experiments 1–3, listeners learned talkers' voices in training phases that alternated (**B**) blocks of passive listening with (**C**) blocks of active practice identifying talkers with feedback. (**D**) Listeners were then tested on their ability to identify the talkers without feedback. (**E**–**G**) In the lexical priming conditions of Experiments 1 and 2, listeners also saw subtitles before and during the talkers' speech that primed them to expect to hear certain words. When listening to Mandarin-speaking talkers, these subtitles led listeners to perceive the intended English gloss of each sentence (albeit with a strong Mandarin accent) as they learned to identify the Mandarin voices from these recordings

in talker identification studies (Perrachione & Wong, 2007; Xie & Myers, 2015a; Zarate et al., 2015; *inter alia*). During familiarization (Fig. 1B), participants heard each of the five talkers say the same sentence in turn while the corresponding avatar and talker number appeared on the screen. Listeners heard each talker say the sentence twice (ten familiarization trials). Next, participants completed a ten-trial block of talker identification practice (Fig. 1C). With all five of the talkers' avatars on the screen, listeners heard each of the talkers saying the same sentence from the preceding familiarization block twice in a random order, and they indicated on each trial which talker they believed was speaking by pressing the corresponding number on a keypad. Participants received corrective feedback indicating whether they had chosen correctly, or who the correct talker was. After ten active practice trials, listeners underwent the next block of passive familiarization with a new sentence, and so on until they had been trained on five sentences. Thus, participants completed a total of 100 trials of training: 50 trials of familiarization with each talker (5 sentences × 5 talkers × 2 repetitions) and 50 trials of active practice identifying the target talkers with feedback.

After training was completed, listeners were tested on their ability to identify the talkers. They again saw all five talkers' avatars on the screen and indicated which of the five speakers they believed said a sentence (Fig. 1D); however, in the test phase, participants did not receive feedback. Participants heard the same sentences during test that they had heard during training. While it is often desirable to test of novel materials to ascertain generalization of talker identity to new speech materials, doing so frequently results in a performance decrement (e.g., McLaughlin et al., 2015; Perrachione & Wong, 2007). To accommodate the possibility that the beneficial effects of the subtitles were small, we chose to use the same sentences during the test phase to maximize listeners' familiarity with the stimuli and thus their potential opportunity to use lexically-derived cues for talker identification. The order of sentences and talkers was randomized, and participants' talker identification abilities were tested in 50 test trials (5 talkers × 5 sentences × 2 repetitions).

For the two conditions where subtitles were used to prime lexical expectations, each subtitle was displayed on the screen two seconds before the presentation of the recording, so that listeners would have enough time to read it and form an expectation about the speech content of the upcoming sentence. Subtitles accompanied the presentation of all speech stimuli in these conditions, including during familiarization, practice with feedback, and at test (Fig. 1 E–G). In the conditions without subtitles, a blank screen appeared for two seconds at the beginning of each trial, such that the timing of these experiments was the same.

Transcription of speech in the foreign-language conditions

To ascertain whether English-language subtitles accompanying the Mandarin speech were effective at eliciting the intended English lexical representations, half of the participants ($N = 16$) undertook an additional sentence transcription task after completing the talker identification test in each Mandarin condition. In this self-paced transcription task, participants heard, in a random order, each of the five talkers saying each of the five sentences from that condition (25 trials). Participants were instructed to "type the sentence exactly as you heard it," and were told they were free to do so however they thought best reflected what they heard. Participants could

see their responses during each trial while typing them; otherwise, no other information (particularly, no subtitles) appeared on the screen during the transcription task.

Transcription of the hybrid sentences during the two Mandarin conditions were scored on a number of dimensions, including (1) whether the sentence exactly matched the intended English gloss, (2) the proportion of words from the intended English gloss that were transcribed as intended, (3) whether the sentence was transcribed using only real English words, and (4) whether the sentence transcription contained any real English words. Sentence transcriptions were assessed conservatively; for instance, if a participant submitted the transcription, "my friends need your jelly," for the target gloss, "my friend needs some jelly," this was assessed to be (1) an incorrect transcription of the target sentence, (2) a correct transcription of 2/5 words, and (3/4) a transcription containing any and all English words.

Data analysis

In this and the subsequent experiments, data were analyzed using (generalized) linear mixed-effects models implemented in the libraries *lme4* (v1.1-21), *lmerTest* (v3.1-0), and *car* (v3.0-2) implemented in *R* (v3.5.3). Significance was based on the criterion $\alpha = 0.05$, with degrees of freedom based on the Satterthwaite approximation of the degrees of freedom.

Results

Talker identification

Talker identification was operationalized as participants' accuracy on each trial of the test phase of each condition. These scores were submitted to a generalized linear mixed model for binomial data. Fixed factors in the model included *language* (English, Mandarin), *subtitles* (no subtitles, with subtitles), and their interaction. The model's random effects structure included by-participant slopes for both fixed-effects terms and their interaction and correlated by-participant intercepts, as well as by-item intercepts for the nested random factors of talker and sentence. The contrast structure specified for the model included deviation coding for both fixed factors. Significance of fitted model terms were assessed using a Type-III ANOVA with Wald chi-square tests. Significant effects were followed by testing the relevant contrast of model terms to ascertain direction and effect size. Participants' talker identification accuracy in each condition is summarized in Table 2 and illustrated in Fig. 2.

The ANOVA on the linear mixed effects model revealed a significant main effect of *language* ($\chi^2(1) = 36.16$, $p \ll 0.0001$), with the corresponding contrast on the linear model revealing significantly better performance in English than in Mandarin ($\beta = 0.99$, $SE = 0.16$, $z = 6.01$, $p \ll 0.0001$). The main effect of *subtitles* was not significant ($\chi^2(1) = 0.06$, $p = 0.81$), nor was the *language* × *subtitles* interaction ($\chi^2(1) = 0.57$, $p = 0.45$). These results indicate that listeners exhibited the classic language-familiarity effect both with and without subtitles, but that the presence of subtitles did not affect listeners' talker identification accuracy in either language.

Table 2 Talker identification accuracy by condition in Experiment 1

Condition	Accuracy ($\bar{x} \pm s$)
English (no subtitles)	75.5% ± 15.4%
English (with subtitles)	74.7% ± 16.8%
Mandarin (no subtitles)	37.8% ± 15.4%
Mandarin (with subtitles)	39.8% ± 17.5%

Sentence transcriptions

Because the subtitle manipulation had no effect on listeners' accuracy in either English or Mandarin, nor any effect on the magnitude of the language-familiarity effect, it was critical to also examine whether the subtitles manipulation was effective at eliciting the intended English interpretation of the Mandarin speech. To demonstrate whether listeners actually heard English speech when listening to the English-Mandarin hybrid sentences, and whether listeners' propensity to hear the speech as English differed depending on whether it had been paired with subtitles, we measured how many English words listeners used during transcription of those sentences in each Mandarin condition.

The four dependent measures of transcription accuracy described in Table 3 were analyzed in separate linear mixed models. (These were generalized linear mixed effects models for binomial data for the measures of (1) whether the target sentence was transcribed exactly as intended, (2) whether it was transcribed with any English words, and (3) whether it was transcribed with only English words.) The fixed factor in all models was *condition* (no subtitles, with subtitles). The models' random effects structures included by-participant slopes for the fixed effect term correlated by-participant intercepts, as well as intercepts for the random factors of talker and sentence. The contrast structure specified for the model included deviation coding for the fixed factor. The effect of condition was assessed by testing the contrast of that model term to ascertain significance, direction, and effect size.

When listeners had learned Mandarin talkers with accompanying subtitles, they were significantly more likely to provide transcriptions of that speech that were comprised of, at least in part, English words ($\beta = 5.67$, $SE = 2.18$, $z = 2.60$, $p < 0.01$), and were significantly more likely to provide transcriptions that were comprised of only English words ($\beta = 6.99$, $SE = 1.47$, $z = 4.76$, $p \ll 0.0001$) than the condition without subtitles. Furthermore, listeners who had heard the speech with

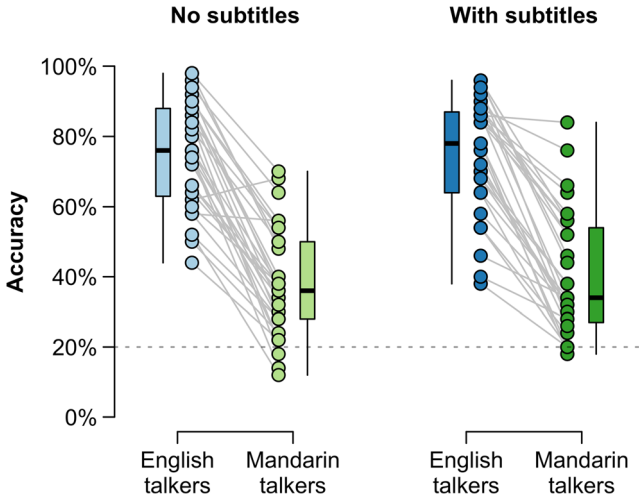

Fig. 2 Talker identification accuracy across conditions in Experiment 1. Talker identification accuracy was significantly and consistently better in listeners' native language (English) than the foreign language (Mandarin) in both subtitles conditions, with no difference in the magnitude of the language-familiarity effect resulting from the subtitles manipulation. *Legend*: Points show accuracy for each participant in each condition; lines connect pairs of points obtained from the same participant. Boxplots show the median (dark line), middle 50% (shaded region) and range (whiskers) of the distribution in each condition. The dashed horizontal line indicates chance (20%)

subtitles were also significantly more likely to hear words from the intended English gloss (β = 0.29, SE = 0.033, t = 8.73, $p \ll 0.0001$) and more likely to give exactly the English gloss intended for each sentence ($\beta = 6.63$, $SE = 3.41$, $z = 1.95$, $p = 0.051$).

Representative examples of listeners' transcriptions of Mandarin sentences that had been presented with/without subtitles are provided in Table 4. Qualitative reports by participants after the experiment also indicated that they reliably believed they were listening to heavily Mandarin-accented English during the Mandarin-with-subtitles condition, and to actual Mandarin speech during the Mandarin-without-subtitles condition.

Discussion

Several prior studies have shown that the presence of familiar words in speech facilitates talker identification (e.g., Bricker & Pruzansky, 1966; Goggin et al., 1991; McLaughlin et al., 2015; Perrachione et al., 2015; Pollack, Pickett, & Sumby, 1954; Xie & Myers, 2015b). Numerous other studies have shown that lexical expectations, including those imparted via priming, are effective at inducing lexical percepts, even from highly distorted speech (e.g., Davis et al., 2005; Ganong, 1980; Holdgraf et al., 2016; Sohoglu & Davis, 2016). Correspondingly, we had hypothesized that, by providing English subtitles during talker identification training in Mandarin, listeners' expectations about the speech would allow them to parse the speech stream into native-language lexical representations, tap into the processes that facilitate talker identification from familiar words, and thereby improve their ability to learn to identify Mandarin-speaking voices compared to when no subtitles were present.

As in previous studies, listeners demonstrated the language-familiarity effect, in that they were better able to learn to identify talkers in their native language than in a foreign language. Additionally, the subtitle manipulation appeared to be effective at inducing listeners to perceive English words from the Mandarin speech. Listeners reported hearing Mandarin-accented English in the Mandarin-with-subtitles condition. They also demonstrated a significant proclivity to use English to transcribe the English-Mandarin hybrid sentences from the subtitles condition, but not when they believed they were hearing Mandarin.

However, the subtitles manipulation had no effect on listeners' ability to learn to identify voices. Listeners did not perform any better in the foreign-language condition when they had the perceptual experience, based on top-down expectations, that the speech they were hearing contained familiar words. This result suggests that, contrary to our hypothesis, familiar words do not afford listeners additional information about, or the ability to form richer memories of, talker identity in the absence of familiar sound patterns.

However, before committing to the theoretical conclusion that familiar words only facilitate talker identification in the presence of familiar sounds (i.e., when listening to native speech), some methodological considerations warrant further exploration. Results from previous studies have indicated that more extensive training may be necessary for listeners to gain advantage of language-specific representations during talker identification in a less-familiar language. When asking bilingual listeners to learn to identify talkers in their native and

Table 3 Use of English in transcription of English-Mandarin hybrid sentences ($\bar{x} \pm s$)

Use of English	No subtitles	With subtitles
Target English sentence transcribed exactly as intended	0.8% ± 3.0%	49.9% ± 38.0%
Proportion of target English words transcribed as intended	18.6% ± 18.2%	78.4% ± 22.7%
Sentence transcriptions using *only* English words	22.1% ± 34.8%	67.0% ± 40.8%
Sentence transcriptions using *any* English words	53.3% ± 32.9%	96.9% ± 8.0%

Table 4 Example transcriptions of English-Mandarin hybrid sentences

Mandarin sentence and target English gloss	Example transcriptions by condition	
	No subtitles	With subtitles
我们看到了小平 wo mən kʰɐn tauʊ lə ɕiau pʰiŋ "Women can do the shopping."	"woman kandalo shalpin" "wome kinda lasa ping" "woman kah da shala ping"	"women can do the shopping" "women can dollar shopping" "women can do all the shopping"
我的有肉丝 wo tə jou ʐou sɨ "Water your roses."	"wu de ye ro su" "what do you roll sue" "wodaya rose"	"water yer roset" "water your roses" "water your rose"
我爱买白松鼠 wo ai mai pai soŋ ʂu "Why, I might buy some shoes!"	"whyamibangshonshur" "wuay me ben shi shu" "waima ba son zu"	"why I might buy some shoe" "why I may buy some shoes" "why I might buy sung shoe"

second language, they exhibit the language-familiarity effect in favor of their native language on the first day of training, but the magnitude of this difference attenuates and eventually disappears after additional days of training (Perrachione & Wong, 2007). In this way, it may be the case that providing additional days of training with the presence of subtitles to prime lexical expectations during foreign-language talker identification training may allow listeners to overcome their unfamiliarity with the sound structure and take advantage of the additional linguistic information source. Second, in Experiment 1, participants were tested only on trained sentences. Many other studies of talker identification have also tested untrained speech stimuli to assess generalization of talker identity knowledge. It may be the case that the information sources made available during lexical priming will be differentially beneficial for recognizing talker identity from trained stimuli versus generalizing talker identification to novel stimuli – a condition where accuracy typically decreases (McLaughlin et al., 2015, Orena et al., 2015; Perrachione & Wong, 2007). We assessed these questions in Experiment 2.

Experiment 2: Multi-day training of foreign-language talker identification with lexical priming

To test whether accessing familiar lexical representations can confer a benefit during talker identification in a foreign language after additional training, we repeated Experiment 1 in a new group of participants, implementing two key changes: First, participants in Experiment 2 underwent 3 days of talker identity training, as opposed to a single session, to give them additional opportunity to learn to use access to word-level representations for foreign-language talker identification. Additional training has been shown to attenuate the language-familiarity effect in bilinguals, but not in monolinguals, suggesting that additional experience may be necessary to gain advantage from linguistic representations when speech is less familiar (Perrachione & Wong, 2007). In this study, we hypothesized that monolingual English speakers may require additional exposure to lexically-primed hybrid speech in order to make use of the lexical representations, analogous to bilingual listeners' attenuation of the language-familiarity effect with further training. Second, we included untrained generalization sentences during the test phase, to assess whether lexical access during talker identity learning would confer any differential benefit to familiar versus unfamiliar speech content in the foreign-language condition. The repetition of speech content at test has been shown to have a more beneficial effect in a native language than an unknown foreign language (McLaughlin et al., 2015).

Methods

Participants

A new sample of native speakers of American-English completed this study (N = 18, age 18–27 years, M = 20.5, 14 female). Inclusion and exclusion criteria were the same as Experiment 1, with the additional requirement that participants in Experiment 2 perform with greater than chance accuracy in all conditions on all days. This additional inclusionary criterion was added in order to limit the analysis to participants who were able to successfully learn the voices. Four additional participants completed the study but were excluded due to failure to meet the accuracy criterion. (Ultimately, exclusion or inclusion of these participants did not affect the outcomes of this experiment.) This study was approved and overseen by the Institutional Review Board at Boston University. Participants provided written informed consent and received monetary compensation for their participation. Participants in Experiment 2 did not participate in Experiment 1.

The sample size was determined by the counterbalance of experimental conditions, and it is in line with prior studies of the role of language in talker identification. Data from the one prior study of cross-language talker identification training

across multiple days (Perrachione & Wong, 2007) suggest a condition-by-session interaction effect on the order of η^2_P = 0.336. Correspondingly, with N = 18 we have 85% power to detect a similar effect, and 80% power to detect effect sizes of $\eta^2_P \geq 0.305$.

Stimuli

Participants learned to identify talkers from recordings of sentences in three conditions: English, Mandarin with subtitles to prime a target English gloss, and Mandarin without subtitles. In the English condition, listeners heard phonetically balanced sentences in English, drawn from a previous talker identification study (McLaughlin et al., 2015). In the Mandarin-with-subtitles condition, listeners heard the English-Mandarin hybrid sentences from Experiment 1. In the Mandarin condition, listeners heard sentences drawn from a set of phonetically balanced Mandarin sentences (Fu, Zhu, & Wang, 2011). Sentences from these corpora were selected because they were of similar length (six–eight syllables) and duration as the English-Mandarin hybrid sentences. Five native speakers of American English (age 20–29 years, M = 23.4) produced the recordings in the English condition. The same ten native speakers of Mandarin (age 19–27 years, M = 23) from Experiment 1 produced both the English-Mandarin hybrid sentences (in Mandarin) and the Mandarin sentences from Fu et al. (2011).

Procedure

Participants learned to identify talkers' voices across three sessions of training and testing on consecutive days. Participants learned a different group of voices in each of the three conditions: English (without subtitles), Mandarin-with-subtitles, and Mandarin (without subtitles). The structure of the training paradigm was identical to that in Experiment 1 (Fig. 1). The five sentences used during the familiarization and practice blocks were the same across all 3 days (within condition). During the test phase of each session, participants were asked to identify talkers from both the sentences that they had been hearing during training, as well as five new sentences each day that they had not heard either during training or during a prior testing session. The new sentences were included to assess how well the participants' knowledge of the talkers' voices generalized to untrained sentences, and whether this differed across conditions.

Participants completed all conditions of the experiment in every session. The order of conditions was counterbalanced across participants, but kept the same for each participant across days. The talkers learned in each of the two Mandarin conditions were the same five-talker groupings as in Experiment 1, and were counterbalanced across participants to control for potential item-specific learning differences.

Results

Learning in each language condition

As in Experiment 1, talker identification was operationalized as participants' accuracy on each trial of the test phases of each condition and day. These scores were submitted to a generalized linear mixed model for binomial data. Fixed factors in the model included *language condition* (English, Mandarin-with-subtitles, Mandarin-without-subtitles), *sentence exposure* (trained, novel), and *training day* (1, 2, 3; as a categorical factor), and all two- and three-way interactions. The model's random effects structure included by-participant slopes for all fixed effects terms and correlated by-participant intercepts, as well as by-item intercepts for the nested random factors of talker and sentence. The contrast structure specified for the model included pairwise differences between levels of the *condition* factor (Mandarin vs. Mandarin-with-subtitles; Mandarin-with-subtitles vs. English) and for the *day* factor (1 vs. 2; 2 vs. 3), and deviation coding for the *sentence exposure* factor. Participants' talker identification accuracy in each condition is summarized in Table 5 and illustrated in Fig. 3.

The ANOVA on the linear mixed effects model revealed a significant main effect of *language condition* ($\chi^2(2) = 67.96$, $p \ll 0.0001$). The corresponding contrasts on the linear model revealed no significant difference between performance on Mandarin-with-subtitles and Mandarin alone ($\beta = 0.17$, $SE = 0.17$, $z = 0.98$, $p = 0.33$), but significantly better performance on English than Mandarin-with-subtitles ($\beta = 2.05$, $SE = 0.26$, $z = 7.79$, $p \ll 0.0001$). The main effect of *sentence exposure* was significant ($\chi^2(1) = 17.26$, $p \ll 0.0001$), with better performance on trained than novel sentences ($\beta = 0.18$, $SE = 0.044$, $z = 4.15$, $p \ll 0.0001$). The main effect of *day* was also significant ($\chi^2(2) = 47.22$, $p \ll 0.0001$), with overall performance on day 2 significantly better than day 1 ($\beta = 0.33$, $SE = 0.08$, $z = 4.43$, $p \ll 0.0001$), and with performance on day 3 significantly better than day 2 ($\beta = 0.22$, $SE = 0.087$, $z = 2.53$, $p < 0.02$).

There was a significant *language condition* × *day* interaction ($\chi^2(4) = 18.77$, $p < 0.0009$), such that the magnitude of the difference between English and Mandarin-with-subtitles was larger on day 2 than on day 1 ($\beta = 0.46$, $SE = 0.19$, $z = 2.45$, $p < 0.015$), but did not differ between days 2 and 3 ($\beta = 0.036$, $SE = 0.20$, $z = 0.18$, $p = 0.86$); however, the magnitude of the difference between Mandarin-with-subtitles and Mandarin did not differ between either days 1 and 2 ($\beta = 0.23$, $SE = 0.16$, $z = 1.45$, $p = 0.15$) or days 2 and 3 ($\beta = -0.016$, $SE = 0.16$, $z = -0.10$, $p = 0.92$).

The *language condition* × *sentence exposure* ($\chi^2(2) = 1.10$, $p = 0.58$), *sentence exposure* × *day* ($\chi^2(2) = 0.83$, $p = 0.66$),

Table 5 Talker identification accuracy by condition in Experiment 2

Condition	Accuracy ($\overline{x} \pm s$)		
	Day 1	Day 2	Day 3
English (no subtitles)	76.4% ± 10.7%	85.8% ± 9.8%	88.2% ± 6.7%
Mandarin (with subtitles)	42.0% ± 13.4%	47.4% ± 12.3%	51.9% ± 13.1%
Mandarin (no subtitles)	41.8% ± 11.0%	42.4% ± 11.9%	46.9% ± 14.7%

and three-way ($\chi^2(4) = 2.17$, $p = 0.70$) interactions were all not significant.

Discussion

Effects of language familiarity and lexical representations

In Experiment 1, the subtitles manipulation failed to improve talker identification accuracy in a foreign language. In Experiment 2, we investigated whether this failure could be overcome with additional training – that is, whether participants required additional exposure to an unfamiliar class of voices to be able to take advantage of linguistic representations to enhance their ability to distinguish and remember those voices (cf. Perrachione & Wong, 2007). We also investigated whether the effect of lexical priming via subtitles might emerge in a more subtle manipulation of trained versus untrained sentence content (cf. McLaughlin et al., 2015).

As in Experiment 1, participants did not perform better in a foreign language condition when given the opportunity to map a foreign-language speech stream onto known words, even when provided with additional opportunity to learn to do so. Participants reliably demonstrated the expected language-familiarity effect across conditions and days (Perrachione, 2018), and the magnitude of the language-familiarity effect did not attenuate with additional training on the foreign language voices, consistent with the one prior multi-day training study in this literature (Perrachione & Wong, 2007).

Correspondence to real-world challenges in talker identification

Although the pop culture phenomenon of "mondegreens" is widely known (Liberman, 2007), and although we observed that our subtitles manipulation was largely

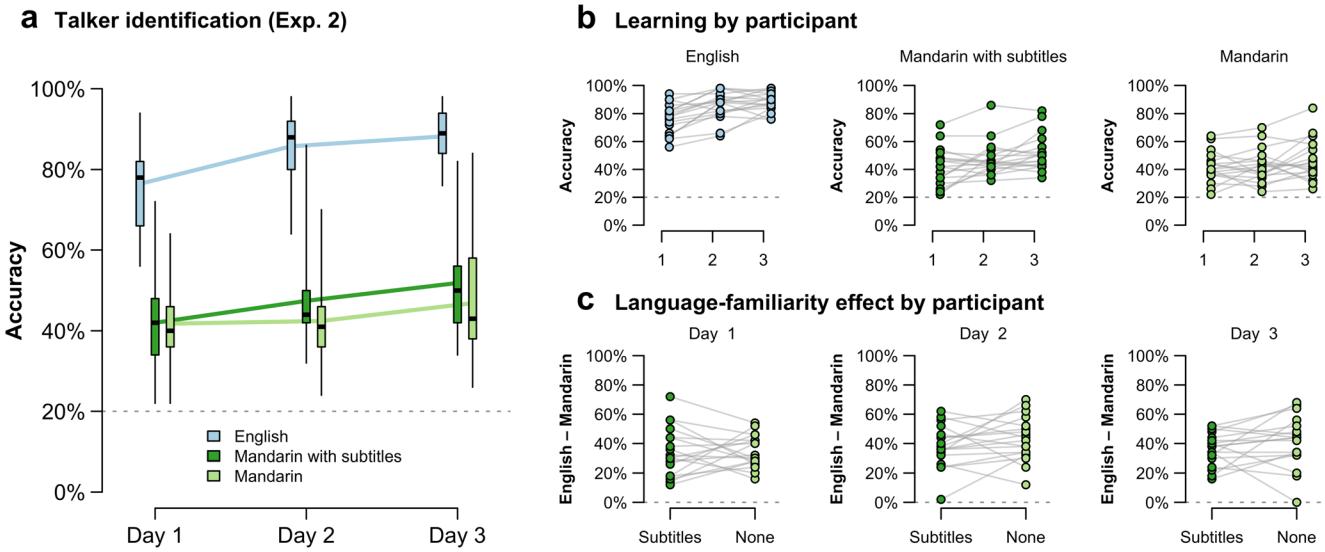

Fig. 3 Talker identification accuracy across conditions and training days in Experiment 2. (**A**) The overall pattern of talker identification accuracy across conditions held constant across all 3 days of training: better accuracy in English, and lower, but on average equal, accuracy in the two Mandarin conditions. On Day 2, the amount of improvement in English exceeded that of Mandarin with subtitles, but the two Mandarin conditions did not differ, and the condition differences remained the same on Day 3. Boxplots show the median (dark line), middle 50% (shaded area) and range (whiskers); lines show change in mean accuracy across days by condition. (**B**) Individual patterns of learning. Points indicate accuracy for each participant, with lines connecting points from the same participant. (**C**) Individual patterns of the language-familiarity effect (talker identification accuracy in English minus accuracy in either Mandarin condition). There was no difference in the Mandarin conditions across days (conventions as in Panel B)

successful in imposing English lexical structure onto foreign language speech (Table 3), this manipulation represents perhaps a rather extreme degree of mismatch between speech phonetics and intended words, particularly given the phonological dissimilarity between Mandarin and English. That listeners were unable to use lexical access to guide talker identification from foreign-language speech raises the question of whether less extreme differences in phonological encoding of known words might also pose a barrier to successfully learning talker identity.

Less extreme than experimentally imposing the sound structure of language onto the words of another is the similar, and more ecological, case of lexical-phonetic mismatch that occurs when native speakers of one language learn to speak another one. Vestiges of speakers' native language persist when speaking in a second language, and these foreign accents come in varying degrees depending on how well language learners acquire the phonology of their second language (e.g., Porretta, Kyröläinen, & Tucker, 2016). Moreover, psycholinguistic research has shown that lexical activation varies as a function of speaker accentedness (Porretta, Tucker, & Järvikivi, 2015).

Although listeners were unable to gain access to additional talker-related information by parsing a speech stream comprised of wholly foreign sound structure into known native-language words in Experiments 1 and 2, we wondered how the *degree* of divergence between known words and familiar sound patterns would affect talker identification. Correspondingly we conducted a third experiment, capitalizing on natural variation in second-language speech proficiency to investigate whether the degree of lexical-phonological mismatch ("accentedness") of talkers' speech affects listeners' talker identification accuracy.

Experiment 3: Identifying talkers with foreign accents

In this experiment, we investigated whether natural variation in the mismatch between word forms and expected phonetic structure affected listeners' ability to learn to identify talkers by voice. Specifically, we trained native English-speaking listeners to learn to identify voices in four conditions that parametrically and ecologically varied the extent to which known words were produced with familiar phonological structure: (1) an *English* condition, with native English-speaking talkers with familiar American accents, (2) a *low-accentedness* condition, with native Mandarin speakers producing English with a slight Mandarin accent, (3) a *high-accentedness* condition, with native Mandarin speakers producing English with a stronger Mandarin accent, and (4) a *Mandarin* condition, with native Mandarin speakers producing Mandarin speech.

The pattern of results in listeners' ability to learn and identify talkers speaking with a foreign accent relative to either native-accented speech or wholly foreign speech will provide further insight into the role of lexical vs. phonetic familiarity in talker identification. If word-level representations play a role in talker identification independent from familiar phonetics, then understanding talkers' speech should play a facilitatory role in talker identification, even if those words are encoded via less-familiar (i.e., foreign-accented) phonetics. Furthermore, so long as the speech is comprehensible, talker identification accuracy should remain high, even as the degree of accent increases. However, if familiar phonetics serves as a "gatekeeper" to the facilitatory role of familiar words in talker identification, then foreign accents should be detrimental to talker identification accuracy, even if speakers are highly intelligible.

Methods

Participants

Native speakers of American-English completed this study (N = 24, 19 female, five male; age 19–28 years, M = 21.8). The inclusion and exclusion criteria were the same as in Experiment 1. This study was approved and overseen by the Institutional Review Board at Boston University. Participants provided written informed consent and received monetary compensation for their participation. Participants in Experiment 3 did not participate in Experiment 1 or 2.

The sample size was determined by the counterbalance of experimental conditions, and is in-line with prior studies of the role of language and accent in talker identification. Based on our previous study of the effects of familiar dialects on talker identification (Perrachione, Chiao, & Wong, 2010), we can expect accent to affect talker identification abilities on the order of d = 0.56. Correspondingly, with N = 24, we have 84% power to detect an effect size of this magnitude, and 80% power to detect effect sizes of $d \geq 0.52$.

Stimuli

Participants learned to identify talkers in four different conditions: (1) *English* spoken by native speakers with American accents, (2) *low-accent* English speech spoken by native Mandarin speakers judged to have the least Mandarin accents, (3) *high-accent* English speech spoken by native Mandarin speakers judged to have the strongest Mandarin accents, and (4) *Mandarin* speech spoken by native Mandarin speakers. In the native-English and accented-English conditions, listeners heard sentences selected from Lists 2, 13, and 22 of the Harvard sentences (IEEE, 1969) spoken by native English-speaking talkers with an American accent (N = 5, age 20–29 years, M = 23.4), native Mandarin speakers whose English

had a light Mandarin accent (N = 5, age 20–26 years, M = 22.8), and native Mandarin speakers whose English had a heavier Mandarin accent (N = 5, age 21–27 years, M = 23.6). In the Mandarin condition, listeners heard phonetically balanced sentences in Mandarin (Fu, Zhu, & Wang, 2011) produced by native Mandarin speakers (N = 5, age 19–24 years, M = 21.6). As in Experiments 1 and 2, all talkers were female.

Accentedness ratings of stimuli

Stimuli for the Mandarin and two Mandarin-accented English conditions were selected from recordings made by 21 Mandarin-English bilingual speakers. We recruited a separate sample of native American-English listeners (N = 12), drawn from the same population as the subsequent talker identification experiment, to rate the degree of accentedness of each talker using principles of comparative judgment (Thurstone, 1927). On each trial, listeners heard recordings of the same English sentence spoken by two native Mandarin speakers. Listeners indicated which of the two talkers they believed had a stronger accent via button press. All possible combinations of talker pairs in both orders were presented, for a total of 210 trials per listener. The proportion of "more-accented" ratings were calculated for each talker in a pair, converted to a z-score, and averaged across listeners. This procedure produced a talker-accentedness rating which reflected not only the rank-order of accentedness across talkers, but also its degree (Meltzner & Hillman, 2005; Perrachione et al., 2014). Recordings from the five speakers with the lowest z-scores (i.e., those least likely to be selected as the "more accented" talker) were used in the "low-accent" talker identification condition; recordings from the five speakers with the highest z-scores were used in the "high-accent" condition (Fig. 4A). From among the remaining eleven talkers that were not rated as the most or least accented, five were selected at random to be speakers in the Mandarin condition. (Example audio recordings from the low- and high-accented talkers are available as Supplementary Materials.)

Procedure

Talker identification training and testing

Participants completed all four accentedness conditions of this experiment in a single session, with the order counterbalanced across participants. The sentences and talkers were not repeated within or between experimental conditions. Training and testing in each condition followed the same procedure as Experiments 1 (Fig. 1A–D). Participants were told they would be identifying talkers who spoke in English, in English with a Chinese accent, or in Chinese. No subtitles were provided in any condition in Experiment 3. The sentences used in the three English and accented English conditions were counterbalanced across participants, and three different sets of Mandarin sentences were used to match the degree of items-level variance in this condition.

Transcription of speech in English conditions

To ascertain how the intelligibility of English speech was affected by the accentedness of the talkers, an additional sample of participants (N = 6) undertook an additional sentence transcription task after completing the talker identification test in each English condition. The structure of this task was identical to the transcription task in Experiment 1.

Transcription of the sentences spoken in English were scored on two dimensions: (1) whether the transcription exactly matched the target English sentence and (2) the proportion of words from the target English sentence that were transcribed as intended. Sentence transcriptions were assessed conservatively, as in Experiment 1.

Results

Sentence intelligibility

Transcription accuracy for the accented sentences was overall extremely high, particularly at the word level (Table 6), but did decrease as a function of talkers' accentedness. The vast majority of deviations from the canonical transcription consisted of lexical neighbor replacements (e.g., "The large house had hot water taps" transcribed as "the latch house had hot water caps"), the addition of a word not present in the canonical version, or, very rarely, attempts to represent the accent phonetically (e.g., "da large house had haut water taps."). There were no instances of attempts at wholly (or mainly) phonetic transcriptions; the vast majority of sentences were transcribed consistently with the speech content intended by the talker (i.e., the canonical Harvard sentence).

The two dependent measures of transcription accuracy described above were analyzed with linear mixed models. For the measures of whether the target sentence was transcribed exactly as intended, a generalized linear mixed effects models for binomial data was used. The fixed factor in all models was *condition* (English, low-accent, high-accent). The models' random effects structures included by-participant slopes for the fixed effect term with correlated by-participant intercepts, as well as by-item intercepts for the random factors of talker and sentence. The contrast structure specified for the model included pairwise differences between levels of the fixed factor (English vs. low-accent; low-accent vs. high-accent). The effect of condition was assessed by testing the contrast of that model term to ascertain significance, direction, and effect size.

Participants provided more accurate transcriptions of native English speakers' recordings than those of low-

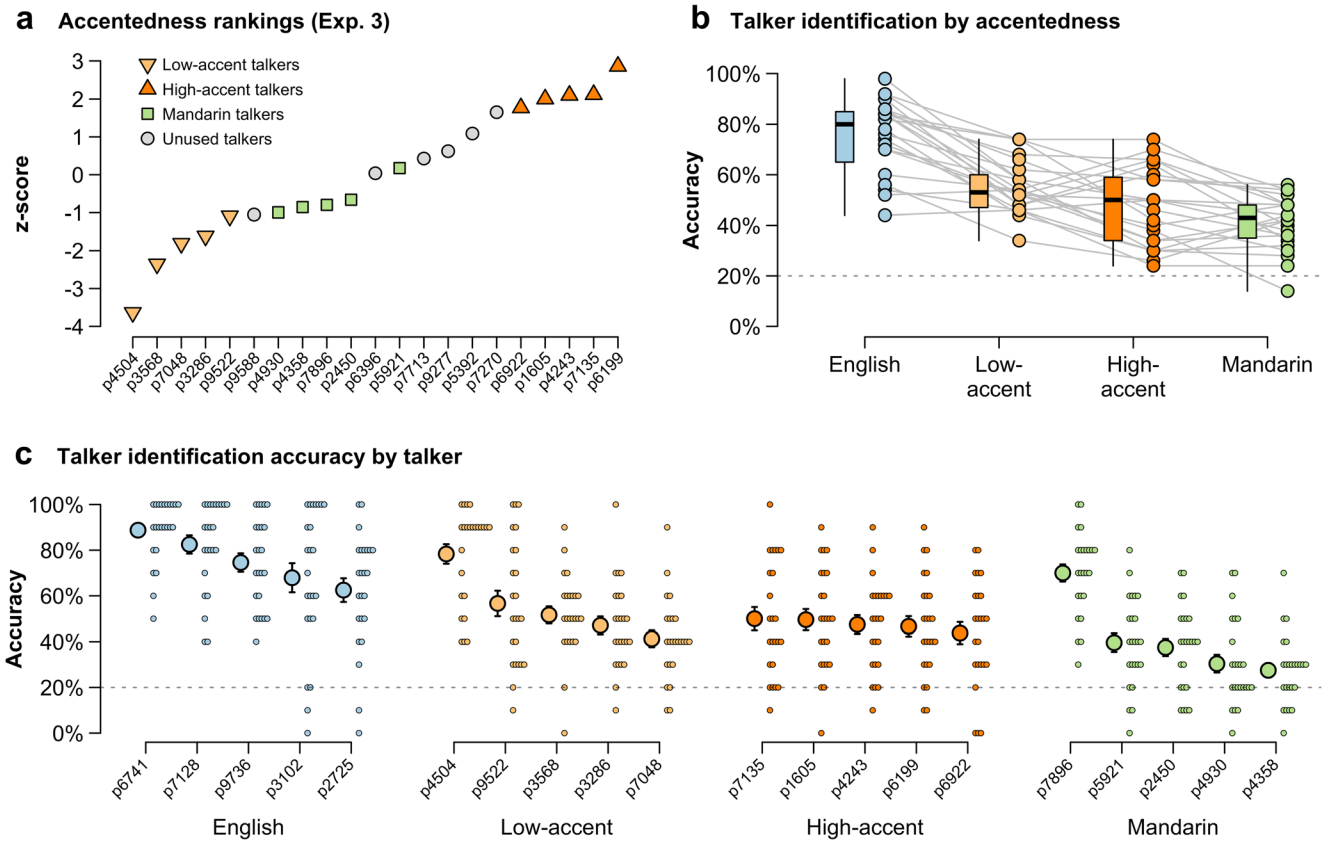

Fig. 4 Results of Experiment 3. (**A**) Listeners' judgments of accentedness for Mandarin-English bilinguals speaking in English. Talkers are ranked by mean accentedness judgments from least to most. Which talkers were selected for use in the low-accent English, high-accent English, and Mandarin conditions are indicated. (**B**) Talker identification accuracy across accent conditions. Plotting conventions as in Figure 2. There was, on average, a monotonic reduction in talker identification accuracy with increasing divergence from the sound structure familiar to listeners. (**C**) Talker identification accuracy across all talkers in each accent condition. Large points show mean identification accuracy for that talker (± SEM) across participants. Smaller points, adjusted along the abscissa to avoid overlap, show accuracy for individual participants

accented talkers (*whole sentence*: $\beta = 2.75$, $SE = 0.66$, $z = 4.17$, $p \ll 0.0001$; *proportion of words*: $\beta = 0.037$, $SE = 0.015$, $t = 2.44$, $p < 0.032$), and provided more accurate transcriptions of low-accented talkers than high-accented ones (*whole sentence*: $\beta = 1.11$, $SE = 0.21$, $z = 5.24$, $p \ll 0.0001$; *proportion of words*: $\beta = 0.047$, $SE = 0.014$, $t = 3.26$, $p < 0.009$).

Talker identification

The dependent measure was participants' accuracy on each trial during the test phases of each condition. These scores were submitted to a generalized linear mixed model for binomial data as in Experiments 1 and 2. Fixed factors in the model included *condition* (English, low Mandarin-accented English, high Mandarin-accented English, and Mandarin speech), *sentence exposure* (trained and novel), and their interaction. The model's random effects structure included by-participant slopes for the fixed effects terms and their interaction and correlated by-participant intercepts, as well as intercepts for the nested random factors of talker and sentence. The contrast structure specified for the model included pairwise differences between levels of the *condition* factor (Mandarin vs. high-accent; high-accent vs. low-accent; low-accent vs. English)

Table 6 Transcription of sentences in Experiment 3 ($\bar{x} \pm s$)

Transcription metric	Condition		
	American English	Less accented	More accented
Target sentence transcribed exactly as intended	93.6% ± 6.5%	75.8% ± 25.5%	61.3% ± 17.0%
Proportion of target English words transcribed as intended	99.1% ± 0.9%	95.5% ± 4.5%	90.8% ± 6.3%

and deviation coding for the *sentence exposure* factor. Significance of fitted model terms were assessed using a Type-III ANOVA with Wald chi-square tests. Significant effects were followed by testing the relevant contrast of model terms. Participants' mean talker identification accuracy in each of the conditions is shown in Fig. 4B and Table 7.

The ANOVA revealed a significant main effect of *condition* ($\chi^2(3) = 24.20$, $p \ll 0.0001$); the corresponding contrasts on the linear model revealed significantly better performance in the native English than the low-accent condition ($\beta = 1.15$, $SE = 0.38$, $z = 3.01$, $p < 0.003$); however, performance did not differ between the low- and high-accent conditions ($\beta = 0.35$, $SE = 0.37$, $z = 0.96$, $p = 0.34$) or between the high-accent and Mandarin conditions ($\beta = 0.29$, $SE = 0.36$, $z = 0.81$, $p = 0.42$). The ANOVA also revealed a significant main effect of *sentence exposure* ($\chi^2(1) = 38.31$, $p \ll 0.0001$), with the corresponding contrast on the linear model indicating that performance was overall higher for trained than novel sentences ($\beta = 0.24$, $SE = 0.04$, $z = 6.19$, $p \ll 0.0001$). The *condition* × *sentence exposure* interaction was not significant ($\chi^2(3) = 1.69$, $p = 0.64$).

Visual inspection of the data in Fig. 4B and Table 7, however, suggests a pattern of monotonically decreasing talker identification as a function of increasing deviation from native English-accented speech. A more granular investigation of the data at the level of participants' talker identification accuracy for each individual talker reveals that talker-level performance is a source of substantial variance within each level of the fixed factor *condition* (Fig. 4C). While modeling talker as a random effect is intended to account for this sort of variance, Experiment 3 differs from Experiments 1 and 2 in that recordings from only five talkers were available per condition (cf. $n = 10$ in the earlier experiments). Because fitting a random effect assumes the levels of that factor are sufficiently well-sampled to obtain estimates of population-level variance, models (and data) with under-sampled random factors may over-estimate the variance related to that factor. When random factors are nested within fixed-factors, as in talkers of various accentedness, this may result in Type II error in estimating the fixed effect.

Thus, with the caveat that we acknowledge that our dataset provided for too few talkers within each level of the fixed factor, we re-ran the model, but without by-talker random intercepts. The ANOVA on this model again revealed a significant effect of *condition* ($\chi^2(3) = 113.64$, $p \ll 0.0001$). The corresponding contrasts on the linear model revealed significantly better performance in the native English than the low-accent condition ($\beta = 1.07$, $SE = 0.15$, $z = 7.15$, $p \ll 0.0001$), in the low- than the high-accent condition ($\beta = 0.32$, $SE = 0.12$, $z = 2.69$, $p < 0.008$), and in the high-accent condition than in Mandarin ($\beta = 0.27$, $SE = 0.11$, $z = 2.42$, $p < 0.02$). (Note that the effect sizes are similar to the previous model, but the error terms are reduced.) The ANOVA again revealed the significant main effect of *sentence exposure* ($\chi^2(1) = 32.79$, $p \ll 0.0001$), and the linear model contrast showing higher performance for trained versus novel sentences ($\beta = 0.22$, $SE = 0.04$, $z = 5.73$, $p \ll 0.0001$). The *condition* × *sentence exposure* interaction again was not significant ($\chi^2(3) = 2.00$, $p = 0.57$).

Table 7 Talker identification accuracy by condition in Experiment 3

Condition	Accuracy ($\overline{x} \pm s$)
English speech (native accent)	75.3% ± 14.8%
English speech (low Mandarin accent)	55.0% ± 10.4%
English speech (high Mandarin accent)	47.5% ± 14.8%
Mandarin speech	41.0% ± 10.3%

Discussion

Participants again reliably demonstrated the language-familiarity effect, with better talker identification in their native language (English) compared to the foreign one (Mandarin). However, listeners' ability to accurately identify talkers speaking English diminished as a function of amount of foreign accent expressed by the talkers, from none to slight to stronger. Depending on how these data are modeled, increasing accentedness either results in a monotonic decrease in talker identification accuracy, or in a decrement in accuracy not significantly different from identifying talkers from foreign-language speech. This result parallels other studies showing reduced talker identification accuracy for speakers of different regional or social dialects (Kerstholt, Jansen, van Amelsvoort, & Broeders, 2006; Perrachione, Chiao, & Wong, 2010; Stevenage, Clarke, & MacNeill, 2012; also cf. Johnson, Bruggeman, & Cutler, 2018), as well as those showing an overall detrimental effect of foreign accent on talker identification accuracy (Doty, 1998; Goggin et al., 1991; Thompson, 1987).

Experiment 3 revealed a number of interesting and new observations into the roles that lexical vs. phonetic familiarity play in talker identification. While previous studies have shown that listeners identify talkers less accurately when they express an unfamiliar accent, none had investigated whether the degree of accentedness had a corresponding, monotonic effect on the extent to which accuracy is reduced. In this experiment, listeners tended to identify more-accented voices less accurately than less-accented voices, paralleling the observation that stronger accents incur greater interference during linguistic processing of speech (e.g., Porretta, Tucker, & Järvikivi, 2015). This also parallels our observation that intelligibility (measured by sentence transcription accuracy) also decreased for each level of accentedness. There is, however, an interesting difference in how these decrements pattern together. Whereas the proportion of words identified correctly

decreased from 99% in the American English accent to 96% in the low Mandarin-accented speech to 91% in the high Mandarin-accented speech, the decrement in talker identification accuracy was much more extreme: from 75% to 54% to 47%, respectively. The decrement in talker identification accuracy for the low-accented voices was particularly remarkable (a reduction of 21%) given the near-ceiling performance in recognizing these talkers' speech (only one in 20 words was heard differently than intended). This suggests that speech intelligibility alone is not the primary driver of listeners' abilities to learn to identify talkers by voice. Instead, the pattern of accuracy decrements appears to more closely correspond to the pattern of exact transcriptions (Table 7), further bolstering the observation that as speech becomes increasingly distorted from listeners' phonetic expectations, talker identification accuracy falls.

It was additionally surprising that listeners' ability to identify talkers from even highly intelligible speech with a stronger Mandarin accent was only modestly greater than their ability to identify talkers speaking a foreign language entirely (47.5% vs. 41%, respectively). Moreover, the decrement in accuracy between the native accent and low foreign accent was much greater than has been observed for voices expressing unfamiliar social and regional dialects (e.g., Perrachione, Chiao, & Wong, 2010), despite the only slight accent expressed by these talkers.

A number of possible mechanisms may explain why the degree of accent should impose an increasing cost on talker identification, even when talkers are speaking in listeners' native language and are highly intelligible. First, as accent increases, so too does the divergence between expected and encountered sound patterns, and listeners may rely primarily on familiar sound patterns during talker identification (e.g., Fleming et al., 2014). Alternatively, as foreign accent increases, the depth of linguistic processing may be reduced (e.g., Porretta, Tucker, & Järvikivi, 2015), which may decrease the extent to which that additional source of information may be available to listeners for talker identification (e.g., Perrachione et al., 2015); however, the high intelligibility scores suggest this account is unlikely. Third, speech perception from an unfamiliar foreign accent is more effortful than from a native accent (Bradlow & Bent, 2008); even though listeners' task was to learn to identify talkers by the sound of their voice, processing the linguistic content of speech is automatic, and the speech perception system may prioritize allocation of cognitive resources for speech comprehension, leaving fewer cognitive resources available for learning talker identity (e.g., Antoniou & Wong, 2015; Bunge, Klingberg, Jacobsen, & Gabrieli, 2000; Heald & Nusbaum, 2014; Kleinschmidt & Jaeger, 2015). Adjudicating between these possible sources of interference will require nuanced experimental designs in future work, but at a summary level the empirical results from Experiment 3 are unequivocal with respect to those from Experiments 1 and 2: Unfamiliar speech sounds impose a cost on learning and identifying talkers by the sound of their voice that supersedes any perceptual or mnemonic benefit gained from hearing talkers say familiar words.

Finally, these data also provide a tangible example of the critical importance of item-level power in psycholinguistic studies. Classical analysis methods applied to these data (e.g., repeated-measures ANOVAs and paired *t*-tests that aggregate data by participant by condition) reveal convincingly significant results between each level of accentedness. However, contemporary mixed models reveal that a large amount of this variation is actually due to differences in the identifiability of individual talkers within conditions. Although the number of talkers per condition in many studies of talker identification has been on the order of four or five (e.g., Bregman & Creel, 2014; Kadam et al., 2016; Orena et al., 2015; Perrachione et al., 2007, 2011; Perrachione, Pierrehumbert, & Wong, 2009; Zarate et al., 2015), the ambiguity of the present results, especially with respect to the theoretically important distinction of categorical versus continuous effects of accentedness, reveals that future work on talker identification must abandon studies with low item-level power in favor of larger numbers of talkers per condition.

General discussion

In the first two of three experiments, we found that the magnitude of the language-familiarity effect was not reduced even when listeners could effectively and convincingly parse a foreign language speech stream into native language lexical representations via priming with subtitles. In Experiment 1, listeners' performance did not improve as a result of primed lexical representations in either the native or foreign language conditions during a single training session. Likewise, in Experiment 2, even though listeners were given multiple training sessions to learn to draw upon lexical representations as a way to improve their talker identification performance, we observed essentially the same pattern of results as in Experiment 1. Finally, in Experiment 3, we found that highly intelligible, ecologically-accented voices were identified with decreasing accuracy as the degree of accent diverged from that of native speakers, that the decrement in talker identification accuracy was much greater than the corresponding decrement in intelligibility, and that the decrement between native-accented talkers to foreign-accented talkers was much larger than between foreign-accented talkers and actual foreign speech. Taken together, these results suggest that the facilitatory contribution of familiar words to talker identification depends on the availability of familiar sounds. Said another way, hearing familiar words in the absence of familiar sound

patterns is not sufficient to improve talker identification in a manner consistent with the language-familiarity effect.

These results help refine our models of the cognitive and perceptual processes that underlie talker identification. Currently, there is evidence to suggest that speech processing and talker identification are functionally integrated, but it has been unknown at what level of linguistic processing this interaction occurs. Some research has indicated the importance of acoustic-phonetic processing as a basis for improved native-language talker identification (Fleming et al., 2014; Johnson, Westrek, Nazzi, & Cutler, 2011; Orena, Theodore, & Polka, 2015; Zarate et al., 2015). Other research has provided similar evidence in support of a facilitatory role of lexical processing in talker identification (Bricker & Pruzansky, 1966; McLaughlin et al., 2015; Perrachione, Del Tufo, & Gabrieli, 2011; Perrachione et al., 2015; Perrachione & Wong, 2007; Pollack, Pickett, & Sumby, 1954). The present results reveal additional nuance to the role of lexical processing in a more complete model of talker identification – that is, lexical processing only appears to play a facilitatory role in the presence of familiar acoustic-phonetic information. When familiar phonetic features are unavailable, it does not appear that listeners are able to make use of lexical access to facilitate talker identification.

Ultimately, these results suggest that the cognitive processes involved in talker identification are supported by a hierarchy of perceptual cues, each of which is likely to depend on successful processing of the previous level. At the lowest level, listeners extract prelinguistic and relatively invariant information about a talker's voice such as fundamental frequency (f_0) and f_0 range, formant dispersion and vocal tract length, and voice quality (e.g., Latinus et al., 2013). Beyond global acoustic properties, listeners gain additional information from acoustic-phonetic features when such features are familiar due to long-term linguistic experience. Naturally, access to phonetic information depends on successful low-level processing and encoding of the auditory signal. Finally, listeners gain additional information about a talker's identity from processing higher-level linguistic information such as through lexical access and memories for words. However, the present experiments suggest that access to this level of information depends on successfully parsing and representing the prior (acoustic-phonetic) level. In all the previous talker identification experiments that have demonstrated beneficial effects of lexical access, lexical information was manipulated in the presence of familiar acoustic-phonetic and phonological structures (Bricker & Pruzansky, 1966; McLaughlin et al., 2015; Perrachione et al., 2015; Pollack, Pickett, & Sumby, 1954; Xie & Myers, 2015b; Zarate et al., 2015). Although these experiments showed, in various ways, that access to word-level representations can improve listeners' abilities to identify voices, they did not explore whether such facilitation depended on successful processing of a lower-level of information, namely the presence of familiar phonology.

An interesting question related to both these results and prior work showing linguistic facilitation of talker identification is whether linguistic representations – at any level – are playing a facilitatory role during *learning* of talker identity versus *recognition* of known talkers. That is, are listeners able to learn talkers better when they can encode their identity through the lens of familiar linguistic representations? Or is it that when talkers' speech contains familiar linguistic structure, listeners have greater access to the indexical features that underlie talker identification? The present results do not directly adjudicate these two possibilities, but we may turn to other paradigms for some insight. We trained our listeners with a certain number of trials for all conditions, and the differences in learning outcomes may suggest that linguistic representations are thus important for learning talker identity. However, in studies that train listeners to criterion (e.g., Bregman & Creel, 2014; Orena et al., 2015), it is often the case that listeners still exhibit language-based differences in subsequent talker identification tests. This suggests that linguistic representations play a facilitatory role in accessing individuating talker information from speech. Additionally, training on native-language talkers and subsequent testing on those talkers speaking a foreign language does not generalize as well as training on foreign-language talkers and then testing on them speaking listeners' native language (Winters et al., 2008), further suggesting that the language-familiarity effect benefit comes from the ability to recognize a talker from their speech more than the ability to learn their identity during training.

The present results also provide new insight into literature on influences of unfamiliar regional and social accents on talker identification, particularly accented talker identification in listeners' native language. Listeners have consistently been shown to perform worse at identifying talkers speaking with an unfamiliar social or regional accent in native language conditions (Doty, 1998; Goggin et al., 1991; Goldstein et al., 1981; Kerstholt, Jansen, Amelsvoort, & Broeders, 2006; Perrachione, Chiao, & Wong, 2010; Stevenage, Clarke, & MacNeill, 2012; Thompson, 1987). In all these cases, listeners putatively had access to lexical information to some extent, since the linguistic content was familiar. However, while talker identification tends to be poorer in an unfamiliar accent than in a familiar one – likely due to less experience with the characteristic distributions of the phonetic features in the unfamiliar accents – across studies, performance in an unfamiliar accent of a native language has consistently been much better than in a foreign language, where both the linguistic and phonetic features are unfamiliar. In this way, it was unclear whether priming access to familiar words (in the absence of familiar phonology) would nonetheless improve talker identification over a fully foreign language condition, even if listeners' performance still did not reach the level of the native language condition (since voices speaking accented L1 speech are still

much better identified than L2 voices). A principal contribution of the present experiments, therefore, is to show that there is indeed a dependency relationship between familiar words and familiar sounds – the former is only beneficial in the presence of the latter, particularly when the latter is very unfamiliar.

A remaining limitation of these experiments is that they cannot distinguish the deleterious effects that cognitive resource allocation to accented speech perception versus speech-sound unfamiliarity might have on talker identification accuracy. It is possible that less accurate performance in accented speech or a foreign language could result from limitations in the deployment of cognitive resources to the task of talker identification versus attempting to process the linguistic content of speech. Because speech is typically comprehensible, even talker identification in a foreign language may automatically impose a processing cost as listeners attempt to understand the speech, to the detriment of having resources available for learning talker identity. There is some evidence that allocation of cognitive resources towards demanding speech perception can incur a cost on a primary learning task (Antoniou & Wong, 2015; Heald & Nusbaum, 2014; Kleinschmidt & Jaeger, 2015). Unfortunately, there is no extant research to indicate how cognitive resource allocation affects learning talker identity, and future, hypothesis-driven studies are needed to address this question directly. For instance, is talker identification poorer from sentences that are harder to understand, such as those with subject-extracted (vs. object-extracted) relative clauses (Gibson, 1998)? Differences in cognitive load notwithstanding, the present results provide important new insight into how various information sources contribute to talker identification because listeners consistently do not benefit from familiar words in the absence of familiar sounds, even when provided multiple exposures of lexical primes across several days of training.

Conclusions

These three experiments suggest that a more complete model of the cognitive processes involved in talker identification includes both acoustic-phonetic and higher-level linguistic processing. Furthermore, they suggest that there is a hierarchical relationship among these linguistic levels, where the facilitatory effects of lexical access in talker identification depends specifically on the availability of familiar acoustic-phonetic forms.

Acknowledgements We thank Sara Dougherty, Kristina Furbeck, Jessica Tin, Emily Thurston, Lauren Gustainis, Jennifer Golditch, Michelle Lee, Jonathan Mirsky, Andrea Chang, Cassandra Chan, Ja Young Choi, Sudha Arunachalam, and Charles Chang for their assistance. We especially thank Rachel Theodore for thoughtful comments on this manuscript. Research reported in this article was supported by the NIDCD of the National Institutes of Health under award number R03DC014045 (to TP). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Open Practices Statement The full set of stimulus recordings, experimental scripts, participant data, and data analysis scripts are available on our institutional archive, OpenBU, at https://open.bu.edu/handle/2144/16460

References

Antoniou, M. & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *Journal of the Acoustical Society of America*, 138(2), 571-574.

Bradlow, A.R. & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, 130(1), 85–95.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America*, 40(6), 1441-1449.

Bunge, S.A., Klingberg, T., Jacobsen, R.B., & Gabrieli, J.D.E. (2000). A resource model of the neural basis of executive working memory. *Proceedings of the National Academy of Sciences*, 97(7), 3573-3578.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, 134(2), 222.

Doty, N. D. (1998). The influence of nationality on the accuracy of face and voice recognition. *American Journal of Psychology*, 111(2), 191.

Fecher, N. & Johnson, E.K. (2018a). Effects of language experience and task demands on talker recognition by children and adults. *Journal of the Acoustical Society of America, 143*, 2409-2418.

Fecher, N., & Johnson, E. K. (2018b). The native-language benefit for talker identification is robust in 7.5-month-old infants. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(12), 1911-1920.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, 11(38), 13795-13798.

Fu, Q. J., Zhu, M., & Wang, X. (2011). Development and validation of the Mandarin speech perception test. *J. Acoust. Soc. Am.* 129, EL267-EL273.

Furbeck, K.T., Thurston, E.J., Tin, J.A.A., & Perrachione, T.K. (2018). Perceptual similarity judgments of voices: Effects of talker and listener language, vocal source acoustics, and time-reversal. *175th*

Meeting of the Acoustical Society of America (Minneapolis, May 2018).

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6(1), 110.

Getz, L.M. & Toscano, J.C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, doi:https://doi.org/10.1177/0956797619841813

Gibson, E., (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1-76.

Goldstein, A.G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition and memory for accented voices. *Bulletin of the Psychonomic Society*, 17(5), 217-220.

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.

Heald, S.L.M. & Nusbaum, H.C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 8, 35. doi: https://doi.org/10.3389/fnsys.2014.00035

Holdgraf, C.R., De Heer, W., Pasley, B., Rieger, J., Crone, N., Lin, J.J., Knight, R.T., & Theunissen F.E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, 7, 13654.

IEEE. (1969). IEEE recommended practices for speech quality measurements. *IEEE Trans. Audio Electroacoust*. 17, 225-246.

Johnson, E. K., Bruggeman, L., & Cutler, A. (2018). Abstraction and the (misnamed) language familiarity effect. *Cognitive Science*, 42, 633-645.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, 14(5), 1002-1011.

Kadam, M.A., Orena, A.J., Theodore, R.M., & Polka, L. (2016). Reading ability influences native and non-native voice recognition, even for unimpaired readers. *Journal of the Acoustical Society of America – Express Letters*, 139, EL6-EL12.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: Effects of accent, retention and telephone. *Applied Cognitive Psychology*, 20(2), 187-197.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148-203.

Köster, O. & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, 4, 18-28.

Kuhl, P. K. (2011). Who's talking? *Science*, 333(6042), 529-530.

Latinus, M. & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, 2, 175. https://doi.org/10.3389/fpsyg.2011.00175

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, 23(12), 1075-1080.

Liberman, M. (2007). Autour-du-mondegreens: Bunkum unbound. *Language Log*. Retrieved from http://itre.cis.upenn.edu/~myl/languagelog/archives/005100.html

Lisker, L. & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384-422.

McLaughlin, D. E., Dougherty, S. C., Lember, R. A., & Perrachione, T. K. (2015). Episodic memory for words enhances the language familiarity effect in talker identification. *18th International Congress of Phonetic Sciences* Glasgow.

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language, and Hearing Research*, 48, 766–779.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, 47(4), 379-390.

Narayan, C., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. *Cognitive Science*, 41(5), 1361-1376.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, 143, 36-40.

Perrachione, T. K. (2018). Speaker recognition across languages. In S. Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception*. Oxford: Oxford University Press. Available on-line: https://hdl.handle.net/2144/23877

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, 333(6042), 595-595.

Perrachione, T. K., Dougherty, S. C., McLaughlin, D. E., & Lember, R. A. (2015). The effects of speech perception and speech comprehension on talker identification. *18th International Congress of Phonetic Sciences*.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition*, 114(1), 42-55.

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology – Human Perception and Performance, 35*, 1950–1960.

Perrachione, T. K., Stepp, C. E., Hillman, R. E., & Wong, P. C. M. (2014) Talker identification across source mechanisms: Experiments with laryngeal and electrolarynx speech. *Journal of Speech, Language, and Hearing Research*, 57, 1651-1665

Perrachione, T. K., & Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, 45(8), 1899-1910.

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, 26(3), 403-406.

Porretta, V., Kyröläinen, A.-J., Tucker, B.V. (2015). Perceived foreign accentedness: Acoustic distances and lexical properties. *Attention, Perception & Psychophysics*, 77(7), 2438-2451.

Porretta, V., Tucker, B.V., Järvikivi, J. (2016). The influence of gradient foreign accentedness and listener experience on word recognition. *Journal of Phonetics*, 58, 1-21.

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

Samuel, A.G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, 12, 348-351.

Samuel, A.G. & Frost, R. (2015). Lexical support for phonetic perception during nonnative spoken word recognition. *Psychological Bulletin & Review*, 22, 1746-1752.

Sohoglu, E. & Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proceedings of the National Academy of Sciences*, 113(12), E1747-E1756.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, 24(6), 647-653.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, 1(2), 121–131.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychology Review*, 34, 273–286.

Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *Journal of the Acoustical Society of America, 120*(4), 2285–2294.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*(6), 781–790.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America, 123*(6), 4524–4538.

Xie, X., & Myers, E. (2015a). The impact of musical training and tone language experience on talker identification. *Journal of the Acoustical Society of America*, 137(1), 419. doi:https://doi.org/10.1121/1.4904699

Xie, X., & Myers, E.B. (2015b). General language ability predicts talker identification. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) Proceedings of the 37th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, 5.