Cognition 250 (2024) 105866

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Full length article

# Language-general versus language-specific processes in bilingual voice learning

Line Lloy, Khushi Nilesh Patil, Khia A. Johnson, Molly Babel *

*Department of Linguistics, University of British Columbia, Canada*

ARTICLE INFO

*Keywords:*
Voice learning
Language processing
Bilingualism
Speech

ABSTRACT

Language experience confers a benefit to voice learning, a concept described in the literature as the language familiarity effect (LFE). What experiences are necessary for the LFE to be conferred is less clear. We contribute empirically and theoretically to this debate by examining within and across language voice learning with Cantonese-English bilingual voices in a talker-voice association paradigm. Listeners were trained in Cantonese or English and assessed on their abilities to generalize voice learning at test on Cantonese and English utterances. By testing listeners from four language backgrounds – English Monolingual, Cantonese-English Multilingual, Tone Multilingual, and Non-tone Multilingual groups – we assess whether the LFE and group-level differences in voice learning are due to varying abilities (1) in accessing the relative acoustic-phonetic features that distinguish a voice, (2) learning at a given rate, or (3) generalizing learning of talker-voice associations to novel same-language and different-language utterances. The specific four language background groups allow us to investigate the roles of language-specific familiarity, tone language experience, and generic multilingual experience in voice learning. Differences in performance across listener groups shows evidence in support of the LFE and the role of two mechanisms for voice learning: the extraction and association of talker-specific, language-general information that is more robustly generalized across languages, and talker-specific, language-specific information that may be more readily accessible and learnable, but due to its language-specific nature, is less able to be extended to another language.

## 1. Introduction

The human voice serves many purposes, not only being the carrier of spoken linguistic messages, but also providing social, indexical, and emotional messages. Voices act as a kind of auditory face (Belin, Fecteau, & Bédard, 2004), which is a signature of talker identity across utterances and even across languages (Johnson & Babel, 2023). Listeners must carefully attend to the voice to accurately and effectively process all of the information that it offers. Comprehending spoken language and recognizing the identity of a talker are distinct, but connected processes (Creel & Bregman, 2011). The complexities of these connections are evident in listener experience and how knowledge materially affects talker identification or "telling voices together" (e.g., identifying the same talker across different utterances), and talker discrimination or "telling voices apart" (e.g., distinguishing two different talkers as different versus same; Lavan, Burston & Garrido, 2019).

### 1.1. The language familiarity effect

Language experience affects listeners' talker recognition capabilities, in both their ability to tell voices apart and tell voices together. Listeners are better at recognizing talkers in their native language than in a language they do not speak, an effect first noticed by Hollien, Majewski, and Hollien (1974) and later explored as the Language Familiarity Effect (henceforth, LFE; e.g., Goggin, Thompson, Strube, & Simental, 1991; Thompson, 1987).[1] The LFE is most robust in tasks that require a listener to recognize or identify a voice (Bregman & Creel, 2014; Goggin et al., 1991; Johnson, Westrek, Nazzi, & Cutler, 2011; Nygaard & Pisoni, 1998; Perrachione & Wong, 2007; Thompson, 1987), but also evident in discrimination tasks, where listeners are simply assessing whether two voices are from the same talker or not (Levi, 2018; Levi & Schwartz, 2013; Neuhoff, Schott, Kropf, & Neuhoff, 2014; Wester, 2012). For a comprehensive recent review of the LFE, see Perrachione (2018).

* Correspondence to: 2613 West Mall, Vancouver, BC, Canada V6T 1Z4.
*E-mail addresses:* alloy@mail.ubc.ca (L. Lloy), khushi81@student.ubc.ca (K.N. Patil), khia.johnson@gmail.com (K.A. Johnson), molly.babel@ubc.ca (M. Babel).

[1] We use the term *native* language here and elsewhere in the paper when we need to closely follow the extant literature. In our own more controlled usage, we eschew the terms *native language* or *native speaker*, opting for more descriptive terms centered around age of acquisition instead (Cheng et al., 2021).

https://doi.org/10.1016/j.cognition.2024.105866
Received 16 October 2023; Received in revised form 12 June 2024; Accepted 21 June 2024
Available online 5 July 2024

While some kind of experience with a language is required for a language to be familiar, and thus have the LFE manifest, it is not clear what constitutes *familiarity* or where (on what level) within one's linguistic system that familiarity needs to reside. The competing hypotheses as to what accounts for the LFE are known as the phonetic familiarity hypothesis and the linguistic familiarity hypothesis (Perrachione, 2018). The gist of the phonetic familiarity hypothesis is that language exposure without active linguistic competence is sufficient for the LFE to evince. Compelling evidence for this comes from pre-verbal infants (e.g., Johnson et al., 2011) and French-exposed English monolinguals in Montreal exhibiting the LFE (Orena, Theodore, & Polka, 2015), and that typological similarity – providing "familiar enough" features – across languages may provide benefit (Zarate, Tian, Woods, & Poeppel, 2015). The LFE is also present when listeners are presented with phonologically legal non-words (Perrachione, Dougherty, McLaughlin, & Lember, 2015), indicating that knowledge about phonetic and phonological distributions results in processing advantages. Language familiarity benefits disappear when speech is reversed and phonological information is no longer intact (Perrachione et al., 2015; Perrachione, Furbeck, & Thurston, 2019), although some evidence to the contrary exists, suggesting that time-reversed speech may preserve some phonological information (Fleming, Giordano, Caldara, & Belin, 2014).

The linguistic familiarity hypothesis posits that there is an added benefit from higher-level, structured, and generative linguistic knowledge. Evidence in support of this hypothesis comes from voice learning becoming more challenging when a familiar language is less easily parsed (e.g., Goggin et al., 1991) and voice learning becoming easier with greater linguistic proficiency in the target language (e.g., Bregman & Creel, 2014; Perrachione & Wong, 2007). Comprehensible speech provides the best performance for voice learning (Goggin et al., 1991; Perrachione et al., 2015; Xie & Myers, 2015). McLaughlin, Dougherty, Lember, and Perrachione (2015) also demonstrate that episodic lexical access enhances talker recognition, and strengthens the LFE, while lexical repetition in an unknown language provides no benefit. While there is evidence that generative linguistic knowledge may not be *necessary* for successful talker recognition, higher-level linguistic knowledge consistently helps in voice learning. Fecher and Johnson (2021) show the improvements in voice learning across development are specific to a familiar language, and not generic improvements in voice learning skills, suggesting that the continued acquisition of linguistic knowledge provides the infrastructure for the boost. Crucially, however, higher-level linguistic knowledge does not uniquely exert an influence on voice learning. Familiar words and syntactic structure are not sufficient when the phonetics are different. McLaughlin, Carter, Cheng, and Perrachione (2019) provided elegant evidence for the hierarchical nature of the lower-level phonetic input and the higher-level linguistic representations, demonstrating that strong syntactic and word level expectations cannot override unfamiliar phonetic distributions. Further, the L in the LFE is dialect- and experience-specific (Perrachione, Chiao, & Wong, 2010). Several studies have found that LFE is reduced or absent in foreign-accented speech (Kerstholt, Jansen, Van Amelsvoort, & Broeders, 2006; Perrachione et al., 2010; Stevenage, Clarke, & McNeill, 2012; Thompson, 1987). While the phonetic familiarity hypothesis and the linguistic familiarity hypothesis are not incompatible, they are hierarchically ordered such that structural linguistic knowledge only confers an advantage when speech is phonetically familiar.

Unfamiliar or less familiar phonetic distributions are connected to a related observation: the other accent effect (OAE). Evidence that listeners show better voice learning with their own accent compared to another accent is somewhat inconsistent in the literature (e.g., Senior, Hui, & Babel, 2018; Yu, Schertz, & Johnson, 2021). This could be for several reasons. It may be that the OAE is dependent on the social value of the accent, with prestigious accents eliciting more attention than others (Sumner & Kataoka, 2013). Listeners' experiences with the accents may also be variable and hard to quantify. Even within an accent, familiarity and experience matter for voice learning, with better known voices exhibiting benefits in voice processing (Kanber, Lavan, & McGettigan, 2022; Lavan et al., 2019).

### 1.2. Bilingual talkers and listeners

The LFE is built upon the concept that, in the populations being studied, one language is familiar and one is not. Bilingual talkers and listeners who have access to and familiarity with the languages under investigation offer a unique opportunity for assessing the LFE. The term bilingual is generally used in the LFE literature to refer to individuals who have the ability to use at least two languages (Lee & Sidtis, 2017), and it is often recognized that bilinguals' language abilities may differ within and across individuals (e.g. Bregman & Creel, 2014; Xie & Myers, 2015). Representing some of the earliest work on the LFE, Thompson (1987) demonstrated that Spanish-English bilinguals show the LFE with both of their languages.

Crucially, the use of bilingual talkers' speech as stimuli is methodologically advantageous, as this provides consistency in the idiosyncratic talker-specific acoustic structure of the voices used (Johnson & Babel, 2023). The shared acoustic structure across a bilingual's languages is expected to confer a benefit, allowing listeners to generalize unique talker signatures across languages. At the same time, of course, bilingual voices can and do exhibit language-specific and style-specific variation (Lee & Sidtis, 2017), making cross-language talker recognition more challenging than simply generalizing voice learning within a voice to novel same-language utterances. Cross-talker generalization is also more challenging with vocalizations of different types – acted laughter, authentic laughter, and vowel productions, which do not provide listeners with sufficient information from which to generalize speaker identity for unfamiliar or familiar voices (Lavan, Scott, & McGettigan, 2016). The nature of the variability or what is available in the acoustic-auditory signal generally varies across contexts, with more expressive speech being more prone to errors in voice learning than less expressive speech (Lavan et al., 2019).

Several studies have used bilinguals' voices to test listeners' abilities to identify and discriminate talkers across languages (Levi, 2018; Neuhoff et al., 2014; Orena, Polka, & Theodore, 2019; Wester, 2012; Winters, Levi, & Pisoni, 2008; Zarate et al., 2015). However, to our knowledge only three studies use these bilingual voices to test bilingual listeners, and each of the studies uses bilingual listeners of differing levels of familiarity with the experiments' languages. Levi (2018) tests bilingual speakers familiar with only one of the experimental languages (e.g., a French-English bilingual tested on German and English), while Orena et al. (2019) and Bregman and Creel (2014) test bilinguals familiar with both languages (e.g., a French-English bilingual tested on French and English). All three studies compare the performance of bilinguals with monolingual English speakers. Levi (2018) and Orena et al. (2019) trained listeners in only one language, and tested listeners in both. The language pairs in Levi (2018) and Orena et al. (2019) were related Indo-European languages; English and German for Levi and English and French for Orena et al.. While Bregman and Creel (2014) use English and Korean, typologically unrelated languages, all listeners were trained in both languages, and the listener population knew both English and Korean. Therefore, to our knowledge, there are no studies which train bilingual listeners in one language and test their ability to identify voices across typologically unrelated languages. Additionally, we are not aware of any study which compares the performance of bilinguals who are familiar with both experimental languages to bilinguals who are familiar with only one.

Across studies, there is substantial variation in learning rates, accuracy, and confidence in talker identification, with some of the variation accounted for by the nature of an individual's bilingualism. Sequential — as opposed to simultaneous — bilinguals are better at identifying voices in their L1 than in their L2 (Bregman & Creel, 2014; Xie & Myers, 2015). However, early sequential bilinguals (Korean L1) perform better at talker identification in their L2 than late bilinguals (Bregman & Creel, 2014). These age of acquisition effects are gradient, such that the earlier the exposure to the L2, the better the listener performs. Even if not acquired at an early age, listeners with L2 knowledge of

a language perform better than listeners with no knowledge of that language (Koster & Schiller, 1997). Furthermore, language exposure without comprehension, either though extensive in-lab training (Winters et al., 2008) or through passive long-term exposure (Orena et al., 2015), can also lead to a benefit in talker recognition abilities, with those having exposure to the unknown language outperforming those without exposure.

### 1.3. Current study

In a summary of the state of the field with respect to learning voices across languages, Perrachione (2018, 532) notes an unanswered question is whether listeners' abilities to recognize talkers in a native language is due to "enhanced ability to perceive the relevant features that distinguish an individual talker, to learn those features, or to remember them when one encounters a voice again?" In the current work we answer these questions by testing whether listeners from different language backgrounds have ready access to the relevant features (assessing performance in the first training block), learn at different rates (assessing how much training is necessary to meet a performance criterion in training), and generalize learning of talker-voice associations to novel same-language and different-language utterances (assessing performance at test).

We do this using spontaneous speech samples from a corpus of Cantonese-English early bilinguals (Spontaneous Speech in Cantonese and English, henceforth SpiCE, Johnson, Babel, Fong, & Yiu, 2020), which allows the current study to examine listeners' ability to identify the same voice and its concomitant acoustic structure across two typologically unrelated and phonotactically dissimilar languages. In this work, we attempt to balance natural variability and across-voice equitability by using careful selection criteria for extracting stimuli from SpiCE. Bilingual talkers in the SpiCE corpus have consistent voices across languages such that they should have an identifiable 'auditory face' cross-linguistically (Johnson & Babel, 2023). The listeners in this study fall into a number of different categories with respect to language background, permitting us to test multiple hypotheses about who is expected to perform well on voice learning tasks. Cantonese-English bilinguals are expected to show the LFE, as they have knowledge of both Cantonese and English, the languages used in this experiment. We compare the performance of the Cantonese Multilinguals with English Monolinguals and two other multilingual groups: multilingual individuals who speak a tone language (other than Cantonese) and multilingual individuals who do not speak a tone language. This set of language backgrounds allows us to quantify the benefit of being a speaker of a tone language on voice learning (Xie & Myers, 2015) and a generic bilingual advantage (Levi, 2018), but cf. Theodore and Flanagan (2020). Given that age of acquisition can affect performance, all bilinguals included in this study are early bilinguals, operationalized here as individuals who acquired more than one language before the age of five (Bregman & Creel, 2014).

When training is done in Cantonese, given the effects of language familiarity, we expect Cantonese-English bilinguals to learn voices the quickest compared to the other language background groups. If being multilingual provides an advantage for learning to associate voices to talker identities, then multilinguals are expected to learn talker-voice associations more quickly than monolinguals in both Cantonese and English training conditions. Likewise, if knowledge of a tone language provides an advantage for learning to associate voices to talker identities, then Cantonese and other Tone Multilinguals are expected to learn talker-voice associations more quickly than Non-tone Multilinguals and English Monolinguals. If either of these potential benefits are general benefits to learning talker-voice associations, they will appear in both training conditions. If they are benefits which function similarly to language familiarity effects, then they will only appear under the Cantonese training condition. That is, given English is a familiar language to all groups, no group would be predicted to be at an advantage with the English training condition.

At test, participants are expected to better generalize to novel utterances in the same language they were trained in, regardless of whether or not the participant was familiar with the language they were trained in. Cantonese Multilinguals are expected to more accurately identify voices across languages than other groups, given that they are familiar with both test languages. Training in a language that was initially unfamiliar bestows some increased level of familiarity, though not enough to replace more sustained experience. If familiarity through training is sufficient to provide beneficial effects of language familiarity for phonologically dissimilar languages, then listener groups without Cantonese language knowledge who are trained in Cantonese are expected to be more accurate at test overall than their English-trained counterparts. Additionally, if training in a familiar language leads to language-specific learning of talker-voice associations, then for participant groups who do not speak Cantonese, overall test accuracy and cross-language generalization will be better when trained in Cantonese given that their lack of Cantonese knowledge will force them to rely on language-general cues for talker identity. We also expect to see that listeners with certain kinds of bilingual backgrounds will perform better than other listeners; if there is a multilingual advantage, then we expect Tone and Non-tone Multilinguals to outperform Monolinguals. If there is an advantage for knowing tone languages, then Tone Multilinguals will outperform Non-tone Multilinguals.

## 2. Methods

### 2.1. Participants

Participants were either recruited through the participant pool in the Department of Linguistics at the University of British Columbia and compensated with partial course credit or through Prolific and compensated the approximate equivalent of $15 CAD.

A total of 437 individuals participated in the experiment. Data from individuals who were categorized as Cantonese Multilingual (participant pool = 64, Prolific = 4), Tone Multilingual (participant pool = 61, Prolific = 18), Non-tone Multilingual (participant pool = 56, Prolific = 24), and English Monolingual (participant pool = 12, Prolific = 64) were analyzed further. Individuals who did not provide age of acquisition for all of their languages (participant pool = 7, Prolific = 1), multilinguals who identified the age of acquisition of their non-English language at or after the age of five (participant pool = 74, Prolific = 0) or who spoke languages that are challenging to characterize, either because the language has pitch accent[2] or where some dialects have or are developing lexical tone or in cases where participants were imprecise in their language specification[3](participant pool = 52, Prolific = 0) were excluded from analysis (unless they also spoke an easy to categorize tone language, in which case they were considered Tone Multilinguals), leaving us with data from 303 participants. Participants' self-reported genders were 194 = female, 7 = woman, 1 = women, 92 = male, 3 = other, 2 = agender, 2 = nonbinary, 1 = prefer not to say, and 1 = they, and the participants were mostly young adults (Age: M = 22, SD = 4.99, range=18, 57).

Like previous work, the individuals with the multilingual designation are heterogeneous in their language background (Levi, 2018;

[2] Japanese, as a pitch accent language, thus excluded participants' data from being included in the analysis, but two early Spanish-English bilinguals recruited through Prolific with late initial experience with Japanese – ages 25 and 17 – were included as Non-tone Multilinguals.

[3] One participant reported speaking "Taishanese/Cantonese". While these are both languages in the Yue language family, they are considered different languages. This participant was excluded from analysis, and not included in the Cantonese Multilingual group or the Tone Multilingual group.

**Table 1**
Mean self-ratings of English proficiency by language background with standard deviation for in parentheses. Self-ratings were on a scale from 1 (very low) to 10 (perfect).

| Language background | Speaking proficiency | Understanding proficiency |
|---|---|---|
| Cantonese Multilingual | 9.1 (1.4) | 9.2 (1.3) |
| Tone Multilingual | 9.3 (1.2) | 9.4 (1) |
| Non-tone Multilingual | 9.5 (0.7) | 9.7 (0.7) |
| English Monolingual | 9.7 (0.7) | 9.7 (0.7) |

Theodore & Flanagan, 2020). All participants except the English Monolinguals recruited through Prolific and one of the Non-tone Multilinguals recruited through Prolific completed the Language Experience and Proficiency Questionnaire (LEAP-Q, Marian, Blumenfeld, & Kaushanskaya, 2007), which provides extensive language background information. Participants reported experience with a median of 3 languages (Median = 3, Mean = 3.4, SD = 1.27, range = 2,9). Participants were categorized into their language background groups either based on their self-reports on the LEAP-Q or through screening questions on Prolific, which were then confirmed by LEAP-Q responses. Because we did not have hypotheses about the quantity of language experience required for a language familiarity benefit, we included any individual who listed Cantonese in their languages and began their exposure to the language before age five in the Cantonese Multilingual group. Cantonese Multilinguals self-rated their Cantonese understanding proficiency (Median = 7, Mean = 6.99, SD = 2.27, range = 1, 10) as slightly greater than their Cantonese speaking proficiency (Median = 7, Mean = 6.36, SD = 2.48, range = 1, 10). Individuals in the Non-tone Multilingual and Tone Multilingual groups did not report any experience with Cantonese. To be categorized in the Tone or Non-tone Multilingual groups, participants reported early (before age five) acquisition of a tonal or non-tonal language. The tone languages reported included Mandarin (n = 43); Punjabi (n = 17); Vietnamese (n = 15); Hokkien (n = 7); Chinese, unspecified variety (n = 5); Jin Chinese (n = 2); Fukien (n =1); Shanghainese (n=1); and Thai (n=1). Individuals in the Non-tone Multilingual group reported experience with Spanish (n=42); French (n=39); Tagalog (n =9); Arabic (n =7); Hindi (n = 7); Portuguese (n=7); German (n = 6); Russian (n = 6); Farsi (n=4); Italian (n = 3); Ilocano (n = 2); Japanese, see footnote above (n=2); Albanian (n=1); American Sign Language (n =1); Catalan (n = 1); Dari (n=1); Gujarati (n=1); Hebrew (n=1); Hungarian (n=1); Indonesian (n=1); Latin (n=1); Malayalam (n=1); Nepali (n =1); Polish (n=1); Romanian (n=1); Tamil (n=1); and Turkish (n=1).

A note about our sample size. While our Bayesian statistical approach, which we describe in detail below, means we need not be as concerned with statistical power as with frequentist approaches, one still wants a population sample that provides robust and generalizable results. Our goal was to have a minimum of 20 listeners in each test group for each language background. The achieved counts, which fall shy of our goals in one category, are summarized in Table 4, as we must first analyze training performance before ultimately arriving at these numbers.

Participants' language diversity does not entail lower proficiency in English. Participants' self-ratings of speaking and understanding spoken English were uniformly high with mean ratings above nine ("excellent", on the 1–10 self-rating scale). These ratings are summarized in Table 1. The data do not include responses from the 64 English Monolingual participants recruited from Prolific, who were not asked to complete the language background questionnaire and one Non-tone Multilingual from Prolific, who did not complete the questionnaire. Also not represented in these data are one Cantonese Multilingual and one Non-tone Multilingual from our participant pool and one Non-tone Multilingual from Prolific; these three individuals were administered the questionnaire, but did not provide a self-rating on their English proficiency.

### 2.2. Stimuli

The visual materials, shown in Fig. 1, were designed to be similar to those used in Bregman and Creel (2014). The key design elements were that each "face" was a distinct shape and color, and that they present as similarly pleasant/friendly.

Four female talkers were selected from the SpiCE corpus. Given that accent can have an effect on talker recognition (Senior et al., 2018; Yu et al., 2021), an accent judgement experiment was conducted wherein listeners (n = 23) rated the English accents of each talker in the SpiCE corpus on a visual analogue scale (VAS) in terms of how similar their accent is to the local variety of English. The speech samples were of read sentences to ensure that listeners' ratings were based on pronunciation and voice properties, not lexical or syntactic content. The exact rating prompt was to judge how much each speaker sounds like a speaker of English who was born and raised in Canada with "not at all" and "very much" as the endpoints of the VAS. Talkers judged less local sounding were removed from consideration. From the remaining voices, four female speakers were selected based on their consistently high acoustic similarity to themselves (within language) based on probabilistic linear discriminant analysis (PLDA) scores, a common method for automatic speaker identification. The talkers which were most similar in their consistency with themselves across utterances in both languages were selected as insurance that no talker would be more or less difficult to consistently identify. The four selected talkers from the SpiCE corpus were VF21A, VF22A, VF23C, and VF32A. These speakers were all self-identified Asian women of the ages 21, 22, 23, and 32. VF21A, VF22A, and VF23C were simultaneous Cantonese and English bilinguals, exposed to both languages since birth. VF32A reported exposure to Cantonese since birth and English since age three. VF22A and VF32A identify as from Canada, VF21A identifies as from Canada and Hong Kong, and VF23C identifies as from Hong Kong. All speakers report daily usage of Cantonese and English with different social groups (e.g., parents, significant others, friends, etc.) From the four selected talkers, ten stimuli were chosen from the conversational interview portion of the corpus. Because the stimuli were extracted from an interview, each utterance was unique. Possible stimuli were created by cutting interviews into 3000 to 4000 ms segments of speech, beginning and ending with a pause, which were then further trimmed to remove silence and filler words at the beginning or end of the phrase that were separated from the phrase by a pause. Stimuli containing long within-phrase pauses, hesitation sounds, laughter, codeswitches, or that were otherwise incomprehensible were excluded. Some stimuli were included where the talker is cut-off at a pause mid-phrase. The number of mid-phrase cut-offs was balanced across talkers and languages. A transcription of the content of each utterance is available in an OSF repository, along with the sound files. The selected utterances were quantitatively assessed with respect to pitch- and duration-based measures to ensure equivalence across languages, as described in the following paragraphs.

To quantify cross-talker and cross-language variability in pitch and speech rate, variance ratios were used following Johnson et al. (2011). Mean values by language and talker are shown in Tables 2 and 3. To estimate F0, the voiced intervals in the stimuli were automatically identified in Praat and F0 was estimated from these voiced intervals with the STRAIGHT algorithm in VoiceSauce (Kawahara, de Cheveigne, & Patterson, 1998; Shue, Keating, Vicenik, & Yu, 2011). F0 values were converted to semitones using the value of each talker's language-specific 5% value as the reference value, and values below the fifth quantile were removed from the analysis. Talkers' English utterances had a slightly higher and more variable pitch in raw Hertz [English: $M = 186, SD = 47$] than their Cantonese utterances [Cantonese: $M = 181$, $SD = 42$] and in semitones [English: $M = 5.6, SD = 3.62$; Cantonese: $M = 5.2, SD = 3.3$]. The difference in variance ratios for both of these pitch-related measures, however, was not significant [f0: $F(4,4) = 1, p = 0.50$; semitones: $F(4,4) = 1, p = 0.51$].

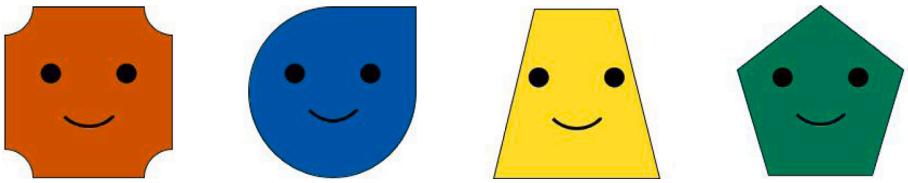

**Fig. 1.** The avatars used in the experiment, depicted as presented to participants.

**Table 2**
Mean f0 in semitones, f0 in raw Hz, duration (in s), and speech rate by language. Standard deviations are in parentheses.

| Language | Semitones | f0 | Duration | Speech rate | Articulation rate |
|---|---|---|---|---|---|
| Cantonese | 5.21 (3.27) | 180.71 (41.66) | 2.84 (0.49) | 3.89 (0.82) | 4.00 (0.80) |
| English | 5.62 (3.57) | 186.04 (46.55) | 2.96 (0.43) | 4.15 (0.67) | 4.32 (0.64) |

**Table 3**
Mean f0 in semitones, f0 in raw Hz, duration (in s), speech rate, and articulation rate for each talker in each language. Standard deviations are in parentheses.

| | Language | Semitones | f0 | Duration | Speech rate | Articulation rate |
|---|---|---|---|---|---|---|
| VF21A | Cantonese | 6.26 (3.13) | 179 (36) | 2.8 (0.37) | 3.962 (0.65) | 3.962 (0.65) |
| VF21A | English | 6.22 (3.99) | 199.05 (51.12) | 2.95 (0.48) | 4.09 (0.55) | 4.18 (0.53) |
| VF22A | Cantonese | 5.69 (3.13) | 161.48 (31.96) | 2.67 (0.56) | 3.35 (0.72) | 3.46 (0.72) |
| VF22A | English | 4.93 (2.92) | 160.04 (29.72) | 2.936 (0.37) | 4.20 (0.57) | 4.24 (0.59) |
| VF23C | Cantonese | 3.49 (2.42) | 167.73 (25.87) | 2.77 (0.60) | 3.85 (0.94) | 3.95 (0.90 |
| VF23C | English | 4.27 (2.71) | 174.57 (29.13) | 2.812 (0.40) | 4.366 (0.64) | 4.71 (0.42) |
| VF32A | Cantonese | 5.45 (3.56) | 211.24 (48.64) | 3.10 (0.35) | 4.387 (0.70) | 4.63 (0.53) |
| VF32A | English | 6.86 (3.78) | 208.12 (51.91) | 3.13 (0.45) | 3.961 (0.90) | 4.17 (0.88) |

Three duration-based measures were made to quantify the amount of phonetic material listeners were exposed to: raw duration, speech rate (number of syllables/second, including pauses), and articulation rate (number of syllables/second excluding pauses). Speech and articulation rate were calculated following de Jong and Wempe (2009). The Cantonese utterances were more variable for duration, speech rate, and articulation rate, but none of these differences reached statistical significance [duration: $F(4,4) = 2.02, p = 0.26$, speech rate: $F(4,4) = 6.12, p = 0.05$, articulation rate: $F(4,4) = 3.4, p = 0.13$].

Stimuli were RMS-amplitude normalized to 65 dB, equalizing the RMS-amplitude for all utterances and voices. However, because participants completed the voice learning task on their own devices, the stimuli were presented at comfortable listening levels as set by each participant.

### 2.3. Procedure

The methods closely follow those of Orena et al. (2019). The experiment was conducted online. Listeners participated on their personal computers and were asked to complete the study in a quiet space and to remove any possible distractions. Listeners were also asked to use headphones, and were required to pass a headphone check before beginning the experiment (Woods, Siegel, Traer, & McDermott, 2017). The experiment consisted of a training and a test phase. In training, listeners learned to identify four voices as they spoke in one language (either Cantonese or English), depending on language condition. They were told that each voice corresponded with one cartoon avatar (see Fig. 1). To facilitate learning, before the training phase began listeners were familiarized with the correct voice-avatar associations. One at a time, listeners were presented with a stimulus from each talker, and an image of the corresponding avatar. Each block in the training phase consisted of 60 randomized trials (four speakers x five sentences x three repetitions) and blocks were repeated until listeners achieved a success rate of 85% or higher within a single block, or until nine training blocks had been completed. If a participant did not attain the target accuracy of 85% in any of the nine blocks, they did not continue on to the test phase. In each trial, listeners were presented with four cartoon avatars in a row, while they heard one of the four talkers speak. Listeners were asked to identify which talker was speaking by clicking on the corresponding avatar. A window of 5000 ms was allowed for listeners to decide which talker they heard, and a gap of 2000 ms occurred between trials. Participants received feedback after each trial as to whether their responses were accurate or not, with an image of the correct avatar appearing on screen.

The test phase was in Cantonese and English, and was identical across all participants. Prior to the beginning of the test phase, listeners were informed that they would now be hearing the same four voices speaking in either English or Cantonese. As in training, during a test trial the listener heard a voice and was presented with an array of all four talkers' avatars. They were then asked to identify the talker by clicking on the avatar's button within a window of 5000 ms. The stimuli used at test were novel stimuli not heard during training. The test phase consisted of one block of 120 trials (four speakers x five sentences x three repetitions x two languages). Test stimuli were presented in random order.

## 3. Analysis

Bayesian multilevel models are used for the analyses using *brms* (Bürkner, 2017) in R using *cmdstanr* on the back end (Gabry & Češnovar, 2021; Stan Development Team, 2021). All of the model results are interpreted following the guidance of Nicenboim and Vasishth (2016) such that when the 95% Credible Interval (CrI) for a given parameter excludes 0, we consider this strong evidence for an effect. The evidence for an effect is described as weak if the CrI includes 0, but the Probability of Direction (PD) is more than 0.95.

All models were fit using four Monte-Carlo Markov chains of 4000 samples each with 1000 warm-up samples per chain. There were no divergent transitions for any model and the $\hat{R}$ values were all < 1.01, suggesting well-mixed chains. Visual inspection of the graphical posterior predictive checks were done for all models, which all indicated that the models fit the data well.

### 3.1. Training

#### 3.1.1. Block 1 performance

Listeners' performance in Block 1 is visualized in Fig. 2, where listeners' mean accuracy is plotted by listener language background

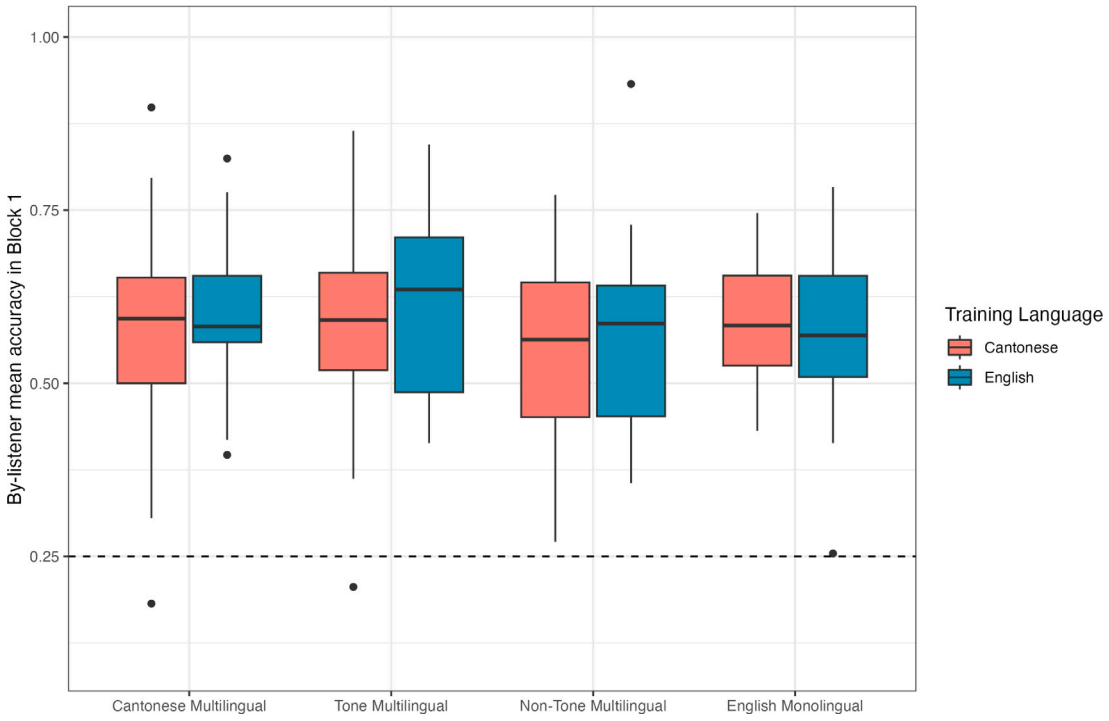


**Fig. 2.** Box-and-whisker plot of by-listener mean accuracy in the first block of the training phase by language background and training language condition.

and training language. The dashed line at 0.25 represents chance performance. Accuracy is generally well above chance – though some individual data points are at or below chance accuracy – but the distributions in the box-and-whisker plots are wide and overlap considerably across groups. These data were analyzed with a Bayesian multilevel regression model using *brms* (Bürkner, 2017) in R using *cmdstanr* on the back end (Gabry & Češnovar, 2021; Stan Development Team, 2021) with a Bernoulli family. Listener accuracy (0, 1) was the dependent variable, and training language (Cantonese, English – treatment coded with Cantonese as the reference level), language background (Cantonese Multilingual, Tone Multilingual, Non-tone Multilingual, and English Monolingual – treatment coded with Cantonese Multilinguals as the reference level), and their interaction were the population-level effects. There were by-participant and by-stimulus group-level random intercepts.[4] Priors were weakly informative priors of normal distributions with a mean of zero and a standard deviation of two for the intercept and one for the population-level parameters (class b). The standard deviations for the group-level effects had a Cauchy distribution with a mean of zero and standard deviation of 2.5 as priors, and correlations used an LKJ prior of concentration two.

The model output for the population-level parameters is summarized in Table A1 of the Supplementary Results. There are no meaningful differences in Block 1 performance. The credible internals all overlap with zero and the probabilities of direction are low (< 0.95).

### 3.1.2. Learning rates for talker-voice association

Within the training phase, we examine how long it takes for listeners to meet the within-block threshold of 85% accuracy across language background groups, at which point participants were able to pass on to the test phase. The empirical data used for this analysis is visualized in Figure A1 of the Supplementary Results. Table 4 reports the numbers of participants in each language background group assigned to each training condition, how many passed into the test phase, and the percent passing rate. It can be gleaned from these summaries of the data that the performance threshold was met much earlier in the English training condition, and that many non-Cantonese speaking participants did not meet the threshold in the Cantonese training condition.

The number of blocks required to meet the performance threshold of 85% is used as a measure of talker-voice association learning rates. To quantify learning rates, we ran a sequential ordinal Bayesian multilevel regression model with the complementary log–log (cloglog) family. Following Bregman and Creel (2014), we make this a conservative assessment, including all of the listeners who fail to meet the threshold in Block 9 of training in the analysis. Training language (Cantonese, English – treatment coded with Cantonese as the reference level), language background (Cantonese Multilingual, Tone Multilingual, Non-tone Multilingual, and English Monolingual – treatment coded with Cantonese Multilinguals as the reference level), and their interaction were the population-level effects. There were by-participant group-level random intercepts.[5] Weakly informative priors of normal distributions with a mean of zero and a standard deviation of 5 were used for the intercept and the population-level parameters (class b).

Because the output of this model is long and complex, interpretation is focused on the visualization of the posterior draws in Fig. 3, which presents the model estimates for the number of blocks for each language group in the two training conditions. Listeners from all language groups met the performance threshold earlier in the English training condition, indicating that the English training condition was easier. For the more challenging Cantonese training, there is strong evidence that Cantonese Multilinguals required fewer training blocks than Non-tone Multilinguals and less strong evidence that the English Monolinguals required more blocks.

[4] Model syntax: Accuracy ~ Training Language * Language Background + (1 | participant) + (1 | stimulus). Because each stimulus is unique to a talker, we could not have talker as a group-level random effect.

[5] Model syntax: Block number ~ Training Language * Language Background + (1 | participant).

**Table 4**
Participant counts by language background group in training and the count of those who passed to the test block.

| Language background | Training condition | *n* in training | *n* in test | Passing rate |
|---|---|---|---|---|
| Cantonese Multilingual | Cantonese | 43 | 35 | 81% |
| Cantonese Multilingual | English | 25 | 24 | 96% |
| Tone Multilingual | Cantonese | 58 | 30 | 52% |
| Tone Multilingual | English | 22 | 22 | 100% |
| Non-tone Multilingual | Cantonese | 48 | 16 | 33% |
| Non-tone Multilingual | English | 31 | 29 | 94% |
| English Monolingual | Cantonese | 47 | 20 | 43% |
| English Monolingual | English | 29 | 27 | 93% |
| TOTAL | Cantonese | 196 | 101 | 52% |
| TOTAL | English | 107 | 102 | 95% |

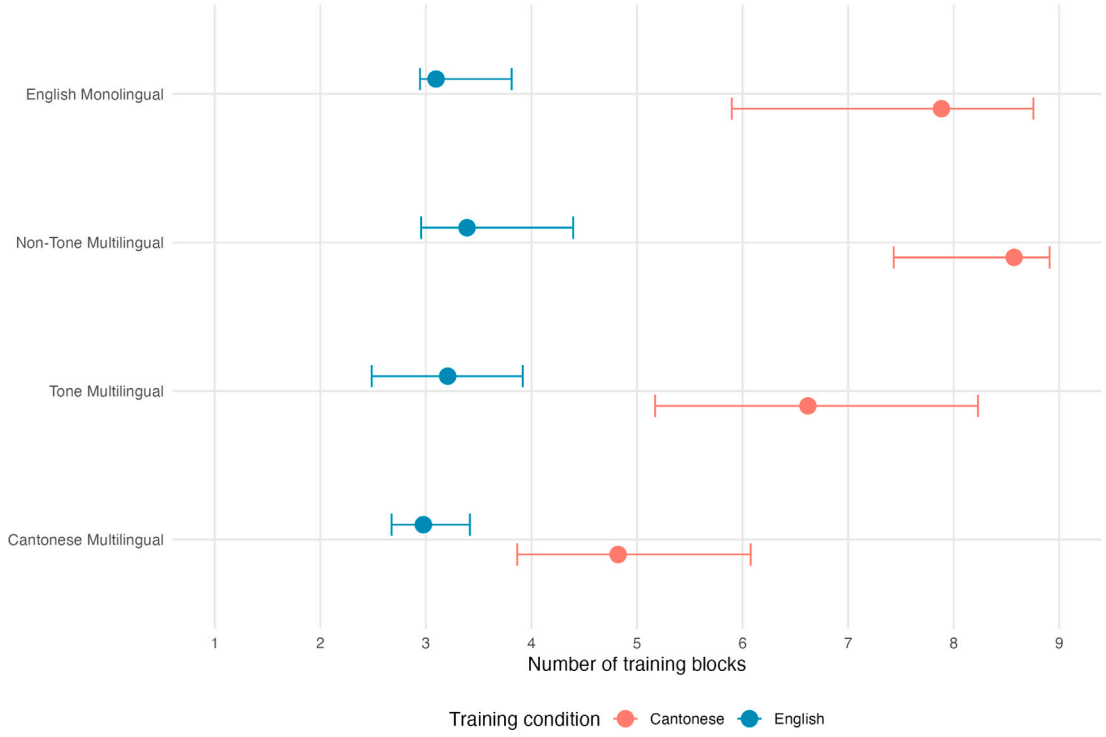


**Fig. 3.** Posterior draws of model predicting the number of training blocks completed before reaching nine training blocks or passing the 85% accuracy threshold, which moved listeners to the test phase. Posterior draws are separated by language background group and training language.

### 3.2. Test phase

From an initial count of 303, 203 listeners passed into the test phase after meeting the performance threshold of 85% correct before or during the ninth block of training. Of those 203, six who did not respond to 10% or more trials in the test phase were excluded from analysis, leaving 197 participants whose performance in the test phase was analyzed. The empirical by-listener mean accuracy, separated by listener group, training language, and test language, is presented as box-and-whisker plot in Figure A2 of the Supplementary Materials.

The analysis of the test data was similar to that of the Block 1 data with a Bayesian multilevel regression model and a Bernoulli family. Listener accuracy (0, 1) was the dependent variable, and training language (Cantonese, English – treatment coded with English as the reference level), test language (Cantonese, English – treatment coded with English as the reference level), language background (Cantonese Multilingual, Tone Multilingual, Non-tone Multilingual, and English Monolingual – treatment coded with Cantonese Multilinguals as the reference level), and their interactions were the population-level effects. There were by-participant and by-stimulus group-level random intercepts, with test language within the participant intercepts.[6] Priors were weakly informative priors of normal distributions with a mean of zero and a standard deviation of two for the intercept and one for the population-level parameters (class b). The standard deviations for the group-level effects had a Cauchy distribution with a mean of zero and standard deviation of 2.5 as priors, and correlations used an LKJ prior of concentration 2.

The model output for the population-level parameters is summarized in Table 5. Given the reference levels, the interpretation of the training and test languages is for the Cantonese Multilinguals. Given the substantial overlap with zero in the credible interval, there is no meaningful effect of training language for Cantonese Multilinguals (CrI: [−0.58, 0.14]; PD: 0.89), and weak evidence for an effect of test language (CrI: [−0.88, 0.01]; PD: 0.97). These main effects are usurped by strong evidence for an interaction between test and training language (CrI: [0.74, 1.36]; PD: 1.0). Cantonese Multilinguals were

[6] Model syntax: Accuracy ~ Training Language * Test Language * Language Background + (1 + test language | participant) + (1 | stimulus). Again, because each stimulus is unique to a talker, talker could not be included as a group-level random effect.

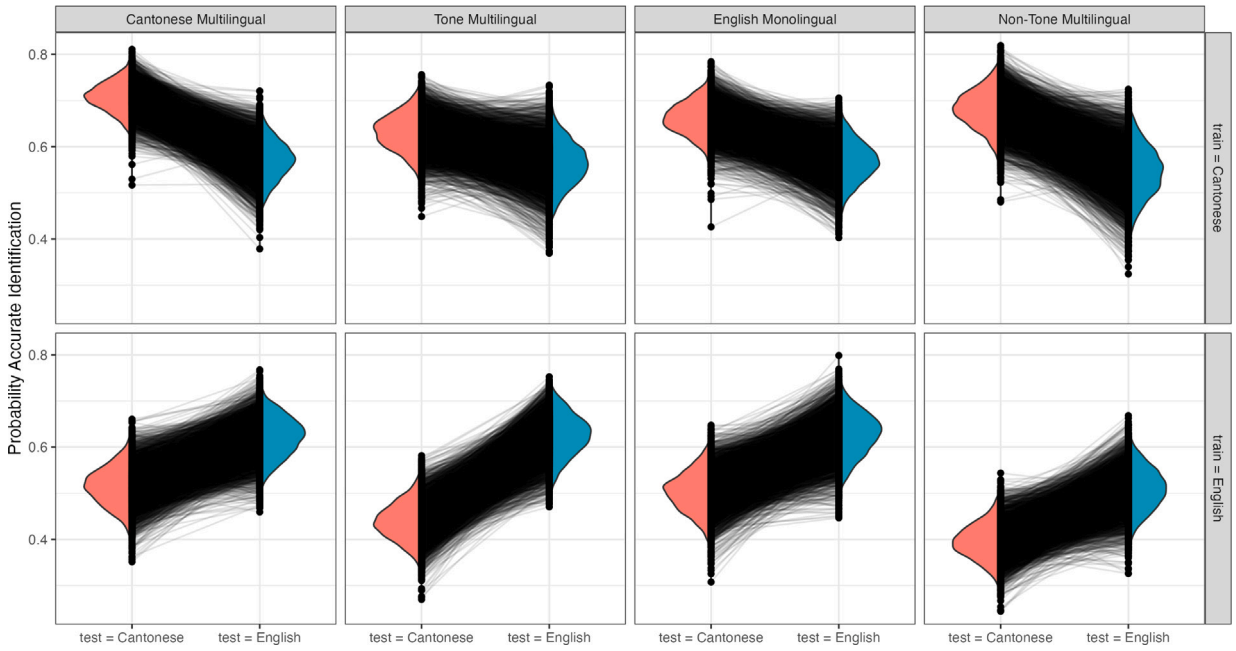


**Fig. 4.** By-listener posterior draws from the model for the test block by language background, training condition, and test language. Lines represent posterior predictions for individual listener performance on Cantonese vs. English at test.

**Table 5**
Population-level or fixed-effect predictors for the Bayesian model for listener accuracy in the test block. The $\hat{\beta}$ estimate, standard error (SE), 95% Credible Interval (CrI), and Probability of Direction (PD) are reported.

| | $\hat{\beta}$ | SE | 95% CrI | PD |
|---|---|---|---|---|
| Intercept | 0.50 | 0.20 | [0.12, 0.87] | 99.38 |
| test language (Cantonese) | −0.44 | 0.23 | [−0.88, 0.01] | 97.20 |
| training language (Cantonese) | −0.22 | 0.18 | [−0.58, 0.14] | 89.08 |
| English Monolingual | 0.01 | 0.20 | [−0.39, 0.41] | 52.33 |
| Non-tone Multilingual | −0.47 | 0.20 | [−0.85, −0.08] | 98.92 |
| Tone Multilingual | 0.02 | 0.21 | [−0.38, 0.43] | 54.10 |
| test language :training language | 1.05 | 0.16 | [0.74, 1.36] | 1.0 |
| test language :English Monolingual | −0.32 | 0.17 | [−0.66, 0.01] | 97.15 |
| test language :Non-tone Multilingual | −0.02 | 0.17 | [−0.35, 0.31] | 55.57 |
| test language :Tone Multilingual | −0.07 | 0.18 | [−0.43, 0.30] | 66.15 |
| training language:English Monolingual | −0.04 | 0.28 | [−0.59, 0.51] | 56.55 |
| training language:Non-tone Multilingual | 0.37 | 0.30 | [−0.21, 0.94] | 89.50 |
| training language:Tone Multilingual | −0.04 | 0.27 | [−0.56, 0.50] | 56.00 |
| test language :training language:English Monolingual | 0.0 | 0.24 | [−0.46, 0.47] | 51.52 |
| test language :training language:Non-tone Multilingual | −0.01 | 0.25 | [−0.51, 0.48] | 50.35 |
| test language :training language:Tone Multilingual | −0.13 | 0.23 | [−0.60, 0.32] | 71.55 |

more accurate at test on the language they were trained in. In fact, all listeners, regardless of language background were more accurate at test with their training language, as there are no meaningful three-way interactions between test language, training language, and language background. This is visualized in Fig. 4, where the model's posterior distributions for language background, training, and test languages are shown with group level performance modeled as distributions, and individual listeners' estimates connected with a line. Listeners trained on Cantonese utterances were more accurate on Cantonese utterances at test, and listeners trained on English utterances were more accurate with the English utterances at test. The generalization of the learned talker-specific traits to novel utterances from a single language was more robust than generalization to a novel language.

The Cantonese Multilinguals were more accurate than the Non-tone Multilinguals when trained in English and tested on English utterances (CrI: [−0.85, −0.08] PD = 98.92), but not better at this training-test combination than the other two language background groups. The evidence in support of the interaction between test language (Cantonese) and English Monolinguals was weak (CrI: [−0.66, 0.01]; PD = 97.51), and indicates that English Monolingual listeners had a more challenging time generalizing English training to Cantonese utterances at test than the Cantonese Multilinguals. A visualization of these posterior distributions can be found in the Supplementary Materials as Figure A3.

To visualize the difference between language backgrounds, Fig. 5 presents the difference in posterior draws between adjacent language background groups. Cantonese Multilinguals are compared to Tone Multilinguals (CntMlt-TMlt); Tone Multilinguals are compared to Non-tone Multilinguals (TMlt-NTMlt); and Non-tone Multilinguals are compared to English Monolinguals (NTMlt-EngMono). If language backgrounds show a large difference in accuracy, then the 95% CrI should not cross zero. This figure neatly illustrates several empirical findings that address theoretical predictions. Examining the comparison of Cantonese Multilinguals and Tone Multilinguals in the first row of Fig. 5, there is no credible difference between groups in terms of their accuracy at test on English items (see the distributions in blue), regardless of their training language, as the 95% CrI for those distributions overlaps with zero. However, there is a credible difference between these two groups in terms of accuracy at test on Cantonese items (see the distributions in red), again regardless of their language training, as the 95% CrI for those distributions does not overlap with

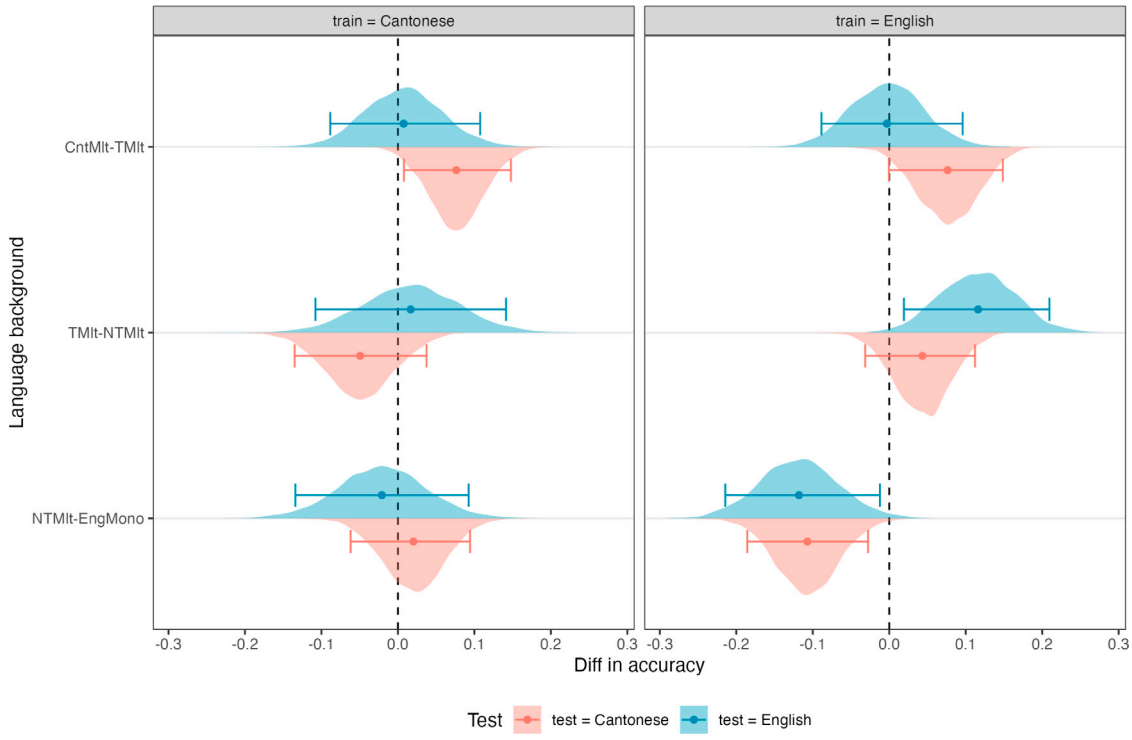


**Fig. 5.** Differences in posterior draws from the model for the test block between adjacent language groups: Cantonese Multilinguals (CntMlt) minus Tone Multilinguals (TMlt); Tmlt minus Non-tone Multilinguals (NTMlt); and NTMlt minus English Monolinguals (EngMono). Posterior distributions are shown by training condition (between-listener) and test language (within-listener).

zero. In fact, Cantonese Multilinguals reliably outperform Tone Multilinguals on Cantonese utterances at test, as this distribution is reliably positive, indicating greater accuracy for Cantonese Multilinguals. These results show that regardless of whether trained in Cantonese or English, Cantonese listeners outperform Tone Multilinguals on the Cantonese utterances at test, while performance on English utterances at test is not reliably different. This illustrates the language familiarity advantage for Cantonese speakers, even when compared to another group which has the possible advantages of bilingualism and knowledge of a tone language. Cantonese listeners apply what is learned with Cantonese training better to Cantonese voices (the classic LFE) and can generalize what was learned from English training to Cantonese utterances better than listeners who speak tone languages other than Cantonese.

Moving on the second comparison in Fig. 5, results show that Tone Multilinguals outperform Non-tone Multilinguals in terms of accuracy for English test utterances when trained in English. This can be construed as evidence for an advantage in talker recognition for tone language listeners, but it only applies to training and testing occur in the familiar language, English.

Finally, examining the last comparison in Fig. 5, we find that counter to predictions, English Monolinguals, when trained and tested with English utterances, had higher accuracy than Non-tone Multilinguals. Despite the evidence in the literature that bilinguals have an advantage in talker recognition, this is perhaps a version of the LFE: English Monolinguals are well-practiced – only-practiced – with English talker recognition, making them more familiar with the experience of recognizing English-speaking talkers.

The more exhaustive way to compare across manipulations is to extract pairwise comparisons across all conditions and language background groups with the emmeans package (Lenth, 2022). In these comparisons, the median point estimate (MPE) is provided on a logit scale and the 95% highest posterior density (HPD). Within language background, across condition comparisons are reported separately from across language background, within condition comparisons. Only meaningful differences are reported in the text, and we refer readers to Tables A2 and A3 in the Supplementary Materials for a full report of all the comparisons.

### 3.2.1. *Across language background, within conditions*

In-test accuracy is measured separately by test item language. This is because the test phase required listeners to generalize the talker-voice associations they learned in training to novel utterances in the same language as training and to a language not presented in training. Generalizing to novel utterances is much more challenging when generalization is across and not within languages. Given this, we examine accuracy for the test items in the same language as training (within-language accuracy), and accuracy for test items in the language not used for training (across-language accuracy). So, for the English language training condition, within-language accuracy is measured in terms of accuracy on English test items, while across-language accuracy is measured in terms of accuracy on Cantonese test items.

In terms of within-language accuracy, when trained in English, a language familiar to all participants, Cantonese Multilinguals (MPE = 0.47, HPD [0.09, 0.86]), Tone Multilinguals (MPE = −0.49, HPD[−0.90, −0.05]), and English Monolinguals (MPE = 0.48, HPD[0.07, 0.87]) were more accurate than the Non-tone Multilinguals. Within the Cantonese training condition, where only the Cantonese Multilinguals were familiar with the training language, Cantonese Multilinguals were better than English Monolingual listeners (MPE = 0.35, HPD[0.04, 0.66]), and there was weak evidence suggesting that Cantonese Multilinguals were better than Tone Multilingual (MPE = 0.22, HPD[−0.07, 0.50]) in the Cantonese training and Cantonese test condition. However, there was no meaningful difference between Cantonese and Non-tone Multilinguals.

In terms of across-language accuracy, within the English training condition, Cantonese (MPE = 0.49, HPD[0.19, 0.79]) and Tone Multilinguals (MPE = −0.44, HPD[−0.75, −0.11]) were both more accurate on Cantonese test items than the Non-tone Multilinguals. Within this condition, there was also evidence that Cantonese Multilinguals were better than English Monolinguals (MPE = 0.31, HPD[0.04, 0.66]).

There were no meaningful differences across language background groups in terms of accuracy on English test items when trained in Cantonese. This suggests that training in Cantonese provides enough experience that the non-Cantonese-speaking participants are able to perform the task well.

Notably, the Non-tone Multilinguals stand out as performing poorly within the English language training condition relative to other groups. When trained in English, Non-tone Multilinguals performed meaningfully worse than all other groups on both English and Cantonese test items. When trained in Cantonese, there were no meaningful differences between the non-tone group and other groups, suggesting that the Non-tone Multilinguals' overall accuracy at test when trained in English (but not Cantonese) was low relative to other groups.

#### 3.2.2. Within language background, across conditions

The by-condition results for each background group are summarized below in terms of the effect of training language on test accuracy (1) by test item language (e.g. the different impact of Cantonese vs. English training on the accuracy of English test items); (2) how well training in a given language generalized to novel utterances languages at test by test item language (e.g. accuracy on English vs. Cantonese items when trained in English); (3) by-condition comparisons of within-language generalization (e.g the impact of training in English on accuracy at English test items vs. the impact of training in Cantonese on accuracy at Cantonese test items); and (4) across-language generalization (e.g the impact of training in English on accuracy at Cantonese test items vs. the impact of training in Cantonese on accuracy at English test items).

In-test accuracy on Cantonese and English test items was differently affected by training condition. With respect to accuracy on English test items, there were no meaningful differences between Cantonese and English training conditions for any group. However, with respect to accuracy on Cantonese test items, all groups were meaningfully more accurate when trained in Cantonese compared to English (Cantonese Multilingual: MPE = −0.83, HPD[−1.13, −0.55]; Tone Multilinguals: MPE = −0.66, HPD[−0.99, −0.35]; Non-tone Multilinguals: MPE = −1.19, HPD[−1.56, −0.84]; English Monolinguals: MPE = −0.79, HPD[−1.09, −0.45]). Taken together, these results suggest that all listeners within each background were similarly accurate on English items at test regardless of training condition, and that regardless of background, those who were trained in Cantonese performed better overall on Cantonese test items relative to those in their same language background who were not. This could mean Cantonese training was generalized more robustly to English such that the Cantonese and English training were equivalent or that generalization to Cantonese test items was only benefited by training in Cantonese, which, for three of the talker groups is because language-general talker-specific information deduced from training in Cantonese only applies to Cantonese.

The Tone Multilingual and English Monolingual groups were meaningfully more accurate on English test items compared to Cantonese test items when trained in English (Tone Multilinguals: MPE = 0.52, HPD[0.06, −1.0]; English Monolinguals: MPE = 0.77, HPD[0.29,1.21]). When trained in Cantonese, on the other hand, Cantonese and Non-tone Multilinguals were found to be meaningfully better at recognizing talkers in Cantonese at test compared to English (Cantonese Multilingual: MPE = −0.60, HPD[−1.04, −0.16]; Non-tone Multilinguals: MPE = −0.58, HPD[−1.09, −0.07]). This complementarity – that Tone Multilinguals and English Monolinguals are better on English test items than Cantonese test items when trained in English and that Cantonese Multilinguals and Non-tone Multilinguals were better at Cantonese test items than English test items when trained in Cantonese – illustrates language background-specific abilities to glean useful and within-language generalizeable information from speech samples.

Nearly all groups demonstrated no meaningful differences in the ability to generalize to the same language at test as they were exposed to in training, regardless of whether that language was English or Cantonese, with the exception of Non-tone Multilinguals, for whom same language generalization was more difficult when training was done in English (MPE = −0.73, HPD[−1.31, −0.17]). When generalizing to a novel language at test, there was again only a meaningful difference for Non-tone Multilinguals, for whom generalizing from English training to Cantonese was more difficult than the inverse (MPE = −0.61, HPD[−1.22, −0.07]). That is, Non-tone Multilinguals appear to learn well from the Cantonese training, despite their lack of Cantonese-specific knowledge.

## 4. General discussion

The myriad results of this experiment simultaneously provide support for the well-documented LFE (e.g. Goggin et al., 1991; Orena et al., 2019; Thompson, 1987), but also provide nuance, identifying where different language groups excel or experience challenges in talker learning and generalization. In a summary of the state of the field with respect to learning voices across languages, Perrachione (2018, 532) notes an unanswered question is whether listeners' abilities to recognize talkers in a native language is due to "enhanced ability to perceive the relevant features that distinguish an individual talker, to learn those features, or to remember them when one encounters a voice again?". In the current work we answer these questions by testing whether listeners from different language backgrounds have ready access to the relevant features (assessing performance in the first training block), learn at different rates (assessing how much training is necessary to meet a performance criterion in training), and generalize what was learned to novel same language and different language utterances (assessing performance at test). We summarize key points in Table 6, and embellish upon these observations below.

### 4.1. Perceive the distinguishing characteristics

Across language background groups or between the Cantonese and English training conditions in the first block of the training phrase, there were no differences in initial performance. One interpretation of this lack of an effect is that talker identification tasks are generally challenging. But, crucially, they are equally challenging for listeners from all language backgrounds. No group is conferred an immediate benefit or disadvantage at perceiving cues to talker identity necessary for distinguishing the four talkers. While answering the question of whether all listeners are equally able to perceive the relevant features to distinguish talkers falls under the purview of a discrimination task, these results nonetheless indicate that all listeners performed above chance, demonstrating that regardless of expertise, be it in the form of language-specific knowledge, knowledge of a tone language, or knowledge of multiple languages, all listeners had equal access to perceive cues of talker identity, but needed time to gain familiarity with individual talkers before they are able build associations, and before language background differences begin to emerge.

### 4.2. Associate the distinguishing characteristics

Listeners' ability to associate talker characteristics to individual talkers was measured in terms of learning rates. The training phase differed in length depending on listener performance; listeners required 85% accuracy with a block (up to a total of nine blocks) to pass to the test phase of the experiment, and as a result, slower learners were required to complete a greater number of training blocks. This method followed Bregman and Creel (2014), which is the only other study to our knowledge which analyses talker-voice association learning rates. Performance across blocks identified differences across language background groups and training language. Listeners from the different language background groups varied in the speed at which they were able to learn the features relevant for distinguishing the talkers. Listeners from all language groups met the performance threshold earlier in the English training condition, indicating that it was easier

**Table 6**
Summary of the key results.

| Main takeaways | |
|---|---|
| Familiarity Effects | Support for the LFE was found; Cantonese-English bilinguals showed the best performance overall, and all groups learned faster in the English condition. |
| Cross-language generalization | Participants were able to generalize talker learning to novel utterances across languages, though accuracy was lower than generalization within language. |
| Differential learning | Training in an unfamiliar language was more beneficial, indicating that listeners learn voices differently depending on language familiarity. |
| Tone Benefit | There was weak evidence in support of the benefit of knowledge of a tone language. |
| Bilingual Benefit | There was not conclusive evidence of a bilingual advantage. Monolinguals were sometimes at an advantage in their performance on English training and test items. |

to associate the talker-specific phonetic variation in English to the four voices in the task. This was the expected result for most listeners, given that English is familiar to all listeners in the study; the LFE predicts this advantage for learning English voices (Bregman & Creel, 2014). That this learning benefit existed for English over Cantonese for the Cantonese Multilingual listener groups could be because of the predominant English language context (e.g., the consent form and task instructions were in English, the institutional language is English). It is also possible that this was not due to situational language dominance, but instead reflects the actual English dominance of the Cantonese Multilingual group. Alternatively, it is also possible that listeners were able to latch on to the English voices more readily because those voices exhibited less variation across utterances. While not significant differences, we did observe that the English utterances showed less within-talker variation for duration, speech rate, and articulation rate. Increased within-talker within-language consistency may have made the English voices easier to learn for all listener groups. Indeed, Lavan et al. (2019) observes that more expressive voices, which contain more variability, are more challenging to learn.

There was an LFE for Cantonese for the Cantonese Multilingual group, which required fewer training blocks in Cantonese to meet the performance threshold compared to the Non-tone Multilinguals and the English Monolinguals. These results are similar to those found by Bregman and Creel (2014), who find that Korean-English bilinguals are faster to learn voices in Korean compared to English Monolinguals. The lack of a meaningful difference between Cantonese and Tone Multilinguals in the number of training blocks required is our first evidence that the Tone Multilinguals might have a generic advantage – that is, not a language-specific one – in talker recognition in an unfamiliar language. The lack of advantage for Non-tone Multilinguals, on the other hand, suggests that perhaps the advantage that Tone Multilinguals seem to have in initial talker-voice association learning may be due to experience with tone (Xie & Myers, 2015) more so than a generic bilingual advantage (Levi, 2018; Theodore & Flanagan, 2020).

Compared to Bregman and Creel (2014), listeners in our study were slower when trained in an unfamiliar language. Bregman and Creel found that it look listeners an average of 4.5 blocks to learn voices in an unfamiliar language, while listeners in our study required more training. Fig. 3's visualization of the model predictions for number of required training blocks indicates that all listener groups, even Cantonese Multilinguals, typically required more than 4.5 training blocks for the Cantonese training condition. One possible explanation for this difference is our use of spontaneous speech samples. The use of spontaneous speech increases the variability within the learning data, increasing within-talker variability and eliminating the possibility of making direct across-talker comparisons in how individual lexical items are produced (Lavan et al., 2019). This added difficulty enriches the learning process while making learning more naturalistic. The added variability posed by spontaneous speech samples also affected the familiar language, English, though not to the same degree. Monolingual and early bilingual listeners in Bregman and Creel (2014) learned English voices within 2.5 blocks on average, while the posteriors for our model predict that listeners of all groups within our study learn voices within 3–4 blocks on average. Altogether, this suggests that while the LFE is robust to the increased variability of spontaneous speech, such natural variation does make the voice learning task more challenging.

### 4.3. Generalize talker identity features

Individuals who passed into the test phase met an accuracy threshold in training ensuring that all listeners successfully learned the four voice-identity pairings in the language they were trained in. This guarantees a relatively high level of knowledge going into the test block. The test utterances were always different from training, thus testing listeners' ability to generalize what was gleaned from training to new utterances. However, while listeners were trained in only one language, they were tested in two languages. Listeners were tested not only on their ability to generalize talker-voice associations to novel utterances, but also to a language novel within the experiment, and in some cases novel to the listeners more generally (i.e., Cantonese for the three groups who do not know Cantonese). Previous literature has demonstrated that listeners rely on language-dependent cues when learning to recognize voices, and are therefore less accurate at identifying talkers in a novel language at test (Orena et al., 2019; Winters et al., 2008). Our study replicates these findings, while adding nuance to our understanding of how listeners with different language backgrounds use language-specific and language-general features to learn talker-voice associations.

Individual listener performance was quite varied – as evidenced by the size of the whiskers in Figure A2 in the Supplementary Materials, and the width of many of the posterior distributions – but at the group level, talker identification accuracy was well above chance, even cross-linguistically. This suggests that listeners can take advantage of voices' cross-linguistic signatures (Johnson & Babel, 2023), which is presumably what allows listeners to accurately identify them even when speaking a language that listeners have not heard them speak before.

Listener background and the training language affected the ease with which listeners were able to reliably extract and associate the talker-specific acoustic detail with unique identities. In the test block, there was clear evidence of language-specific learning; listeners were consistently more accurate in the test block on the language they were trained in, even when that language was unfamiliar and no benefit of the LFE was conferred. This result differs from the findings of Winters et al. (2008), who find that listeners trained in an unfamiliar language generalized that learning fully to their familiar language at test, showing no significant difference in accuracy across languages at test. This may be due to any number of differences in experimental design. Our study used spontaneous speech excerpts, while theirs used consonant-vowel-consonant (CVC) words, and our study used one training session while theirs trained listeners more extensively over two days. These differences made learning in our study more difficult, and perhaps explain why listeners did not fully generalize cross-linguistically.

Within-language generalization to novel utterances was more robust than across-language generalization. This indicates that while an individual bilingual voice shares substantial acoustic structure across languages (Johnson & Babel, 2023) that listeners can ultimately be trained to use (Winters et al., 2008), this is insufficient compared to the benefit bestowed within language, given that some talker attributes are language-specific (Lee & Sidtis, 2017; Orena et al., 2019). Cross-language generalization may be challenging for a number of reasons.

In terms of the acoustic substance, a bilingual's two language codes may simply differ. For example, an individual may have a lower f0 range in Cantonese than English (Johnson & Babel, 2023), challenging a listeners' ability to generalize based on one language's talker-specific associations for pitch. The challenge may also relate to how talker-specific information is encoded. To again use the pitch example, given that Cantonese is a language with lexical tone, an individual speaker's pitch in Cantonese is going to simultaneously function as a linguistic feature and a talker-identity feature.[7] When speaking English, that individual's pitch will primarily be a talker-identity feature. It may be more challenging to generalize across differential uses of pitch cross-linguistically. Additionally, however, some talker-specific information is also language-specific; a talker may have a characteristic VOT in one language but not another, for example. We return to this discussion of language-specific and language-general talker identifying features later within this section.

Importantly, the ability to generalize across languages differed according to language group and training. For Cantonese Multilinguals, training in Cantonese resulted in improved performance on Cantonese test trials compared to English training. The lack of the reverse finding for Cantonese Multilinguals – that English training is not better than Cantonese training for performance on English test trials – suggests that Cantonese Multilinguals are better able to generalize talker-specific/language-general information from Cantonese to English than vice versa. This may relate to the fact that the Cantonese Multilinguals in the study population identify English as their more dominant language. Among participants who are not familiar with Cantonese, training in Cantonese was beneficial. For all groups, performance on English at test was not different across conditions. However, performance on Cantonese test performance was improved by training in Cantonese, with better overall performance when trained in Cantonese.

Additionally, for all groups, listeners were better able to generalize to the novel language at test if they were trained in Cantonese. This suggests not only a familiarity benefit from training in Cantonese, but potentially also a benefit for generalizing to a novel language when training occurs in a language that listeners do not know. These results suggest that if a listener is learning without the guidance of linguistic expertise, then they may rely more on language-general talker information, resulting in learning that is general enough to be broadly applicable for recognizing the same talkers speaking in another language. On the other hand, when learning voices in a known or familiar language, language expertise may improve learning of talker-voice associations, but may result in more language-specific learning and as a result does not robustly generalize to novel languages at test. Winters et al. (2008) found a similar effect with their monosyllabic training stimuli. English listeners trained in German, an unfamiliar language, are better at generalizing voice learning from across languages than listeners trained in their native language (Winters et al., 2008).

Of course, this does not mean a lack of benefit for familiar languages. The results also provide support for the LFE, consistent with other LFE studies (e.g. Bregman & Creel, 2014; Orena et al., 2019; Perrachione & Wong, 2007; Thompson, 1987). Cantonese-English bilinguals on the whole performed better than other groups. Within-language Cantonese testing showed, unsurprisingly, an advantage for the Cantonese Multilinguals compared to the English Monolinguals and the Tone Multilinguals. Both Cantonese and Tone Multilinguals were better able to generalize what was learned from the English training to the Cantonese test items than the Non-tone Multilinguals, and the Cantonese Multilinguals were also better under these conditions than the English Monolinguals.

Training in Cantonese, however, reduced language familiarity effects in some areas.[8] Across-language testing of English test items for listeners trained in Cantonese showed no meaningful differences across language background groups. All listeners were equally able to generalize Cantonese learning to English. This lack of difference suggests that no one group was at an advantage over others when it came to generalizing Cantonese learning to English, demonstrating that training in Cantonese gave participants unfamiliar with Cantonese with enough of a benefit to reduce the presence of LFE , similar to effects in previous literature which found that training in an unknown language reduced LFE effects (Levi, 2018; Winters et al., 2008). Additionally, within language groups, there was no meaningful difference in accuracy at test in terms of within or across language generalization when trained in Cantonese, and accuracy on English test items was equivalent whether trained in Cantonese or English. Taken together, these suggest that for those unfamiliar with Cantonese, their challenging training in Cantonese was sufficient to help them perform similarly to Cantonese Multilinguals.

This suggests that listeners may use different strategies or mechanisms for gleaning talker-specific information from familiar versus unfamiliar languages. Taken together with the results for Cantonese Multilinguals, it may be that training in a non-dominant language likewise relies less upon language-specific information. In other words, when presented with an unfamiliar language, the only listener recourse is to attend primarily to the talker-specific, language general information that is known to be relatively stable across one's languages (Johnson & Babel, 2023). Given this stability, this information can be more easily generalized to another language. When attending to a familiar language, a listener is armed with structured linguistic knowledge that guides the voice learning process, enhancing the relevance of some acoustic dimensions (that are known to be important for contrast in that language) and attenuating sensitivity to others. This renders the voice learning less generalizeable to a language lacking those linguistic structures. So, while access to structural linguistic knowledge may improve a listeners' ability to recognize talkers within a language (Bregman & Creel, 2014; Perrachione et al., 2015; Perrachione & Wong, 2007; Xie & Myers, 2015), these results suggest that it may reduce a listener's ability to generalize talker learning across languages. This finding adds to the discussion on phonetic familiarity versus linguistic familiarity; while both kinds of familiarity may boost talker learning within a language, access to both lower level and higher level language-specific features may hinder cross-language generalization abilities.

The results here provide weak support for the advantage of knowledge of a tone language, as has been proposed previously (Xie & Myers, 2015). Tone Multilinguals outperformed Non-tone Multilinguals and monolinguals in some conditions. However, the tone advantage was not observed across the board. For example, Non-tone Multilinguals were not significantly worse than Cantonese Multilinguals when trained and tested on Cantonese, where Tone Multilingualss were. Likewise, there was not always clear evidence for a multilingual advantage. Non-tone Multilinguals performed particularly poorly when trained and tested on English. This result is somewhat surprising, and contrasts with our prediction that bilingual groups would have a talker learning advantage. A possible interpretation for this may be that some of the Non-tone Multilinguals were faced with an Other Accent Effect (OAE) with the English stimuli. It may be the case that some of the Non-tone Multilinguals from the United States recruited through Prolific would have a different variety of English as their familiar language (e.g., Chicano English, as opposed to the Canadian English samples used in this study.)

[7] This is simplifying the space. An individual's pitch can also index emotional and psychological state.

[8] Worth reiterating is that in order to pass to the test phase, listeners had to demonstrate having learned the voices. In other words, the training in Cantonese was confirmed to accompany learning of the voices.

English Monolinguals, on the other hand, were not at a consistent disadvantage by their monolingualism. As monolinguals, they may have intense expertise in one language, given that their language exposure is nearly exclusive to English. As a result, they may be better equipped to learn associations in English training conditions and generalize those associations within-language at test. Conversely, Non-tone Multilinguals may lack this intense and exclusive English familiarity, while also lacking any benefit conferred by knowledge of a tone language. Our Tone Multilingual listener group was heterogeneous in its language background, but largely composed of Mandarin, Vietnamese, and Punjabi listeners. A tone language advantage may be somewhat gradient in nature. Vietnamese has a tone space that parallels Cantonese in its complexity, contrasting six tones in standard descriptions. Mandarin is typologically related to Cantonese. Punjabi has complex syllable structure and tone, creating more syntagmatic comparisons. Any of these points may confer more or less of an advantage for listeners from different language backgrounds. Future theorizing will benefit from a more granular assessment of the type of tone language experience that might confer voice learning advantages.

## 5. Conclusion

The LFE is present, but not all encompassing. Language familiarity seems to be a robust advantage in the rate of learning, and these data suggest that familiarity affects *what* is learned. Listeners who learn voices in an unfamiliar language appear to take advantage of talker-specific and language-general acoustic characteristics that are generally part of a talker's voice profile in either language (Johnson & Babel, 2023). This results in listeners being better able to generalize from the unfamiliar language (Cantonese) to a familiar language (English). When listeners' voice learning is guided by linguistic knowledge, the learning happens faster, but is less generalizable. What listeners learn about voices in English is less robustly generalized to Cantonese. This does not obliterate a within-language advantage; generalizations within languages is more robust than generalization across. But, ultimately, what is learned from an unknown language is more language-general.

There was evidence that Tone Multilinguals are at a voice processing advantage. This group of listeners learned equivalently quickly as the Cantonese Multilinguals in the Cantonese training phase, and they generalized English training to Cantonese test items as well as the Cantonese-advantaged listeners. While the Tone Multilinguals performed better on the English training and English test than Non-tone Multilinguals, suggesting additional evidence for an advantage for listeners with tone language backgrounds, all listener groups – Tone Multilinguals, Cantonese Multilinguals, and English Monolinguals – all outperformed the Non-tone Multilinguals on English training and test, suggesting that this listener group was at a disadvantage. The Non-tone Multilinguals appear, however, to have an unfamiliar language advantage with generalization; Non-tone Multilinguals and Cantonese Multilinguals performed equivalently well on the training and testing with Cantonese. The results from this study demonstrate that what constitutes an advantage in learning to associate a voice with a particular talker may not translate into an advantage in learning to novel utterances; while tone bilinguals had an advantage in learning, they were less good at generalizing their learning.

Altogether, this suggests that talker identity learning must rely on at least two distinct processes: the extraction and association of talker-specific features independent of linguistic knowledge and the use of talker-specific features that are nested within language-specific representations (Winters et al., 2008). Listeners with relevant language experience can take advantage of both of these channels, though inhibiting language processing when a linguistic category exists is difficult. That is, language processing masks talker processing more than talker processing masks language processing (McGuire & Babel, 2020; Theodore, Blumstein, & Luthra, 2015). This is not to suggest that linguistic processing wholly abstracts away from subphonemic information, but rather there is an asymmetry such that parsing content that is linguistically meaningful to the listener can seemingly block sensitivity to subphonemic details more than the reverse. Ultimately, however, the act of "telling voices together" is supported by both language-general and language-specific processes.

## CRediT authorship contribution statement

**Line Lloy:** Writing – review & editing, Writing – original draft, Software, Methodology. **Khushi Nilesh Patil:** Writing – review & editing. **Khia A. Johnson:** Supervision, Methodology, Conceptualization. **Molly Babel:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

## Data availability

There is a link to the OSF repository containing the data and code.

## Acknowledgments

This work has been supported by grants from the Natural Sciences and Engineering Council of Canada to LL and MB. Thank you to Sabrina Luk and Stephanie Chung for their assistance in preparing materials for the experiment and to Roger Lo for advice on the statistical analyses. All materials for this project, including data and code are available on OSF: https://osf.io/8xyvj/?view_only=b310c7b7a6a94594bdd6580bc3f655b3.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2024.105866.

## References

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. http://dx.doi.org/10.1016/j.tics.2004.01.008.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95. http://dx.doi.org/10.1016/j.cognition.2013.09.010.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, Article 715843.

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204.

de Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. http://dx.doi.org/10.3758/BRM.41.2.385.

Fecher, N., & Johnson, E. K. (2021). Developmental improvements in talker recognition are specific to the native language. *Journal of Experimental Child Psychology*, *202*, 104991–105001.

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*, *111*(38), 13795–13798. http://dx.doi.org/10.1073/pnas.1401383111, URL https://www.pnas.org/content/111/38/13795.

Gabry, J., & Češnovar, R. (2021). Cmdstanr: R interface to CmdStan. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

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. http://dx.doi.org/10.3758/BF03199567.

Hollien, H., Majewski, W., & Hollien, P. A. (1974). Perceptual identification of voices under normal, stress, and disguised speaking conditions. *Journal of the Acoustical Society of America*, *56*(S1), S53. http://dx.doi.org/10.1121/1.1914230.

Johnson, K. A., & Babel, M. (2023). The structure of acoustic voice variation in bilingual speech. *Journal of the Acoustical Society of America*, *153*(6), 3221–3238.

Johnson, K. A., Babel, M., Fong, I., & Yiu, N. (2020). SpiCE: A new open-access corpus of conversational bilingual speech in cantonese and english. In *Proceedings of the 12th language resources and evaluation conference* (pp. 4089–4095). Marseille, France: European Language Resources Association, URL https://www.aclweb.org/anthology/2020.lrec-1.503.

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. http://dx.doi.org/10.1111/j.1467-7687.2011.01052.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2011.01052.x, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2011.01052.x.

Kanber, E., Lavan, N., & McGettigan, C. (2022). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology: General*, *151*(4), 897.

Kawahara, H., de Cheveigne, A., & Patterson, R. D. (1998). An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT-suite. (p. 0659). URL https://www.isca-speech.org/archive/icslp_1998/i98_0659.html.

Kerstholt, J. H., Jansen, N. J. M., Van Amelsvoort, A. G., & Broeders, A. P. A. (2006). Earwitnesses: effects of accent, retention and telephone. *Applied Cognitive Psychology*, *20*(2), 187–197. http://dx.doi.org/10.1002/acp.1175.

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

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*(3), 576–593. http://dx.doi.org/10.1111/bjop.12348, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bjop.12348.

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology*, *72*(9), 2240–2248.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(12), 1604.

Lee, B., & Sidtis, D. V. L. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, *20*(3), 174–185. http://dx.doi.org/10.1080/2050571X.2016.1273572.

Lenth, R. V. (2022). Emmeans: Estimated marginal means, aka least-squares means. URL https://CRAN.R-project.org/package=emmeans, R package version 1.8.1-1.

Levi, S. (2018). Another bilingual advantage? Perception of talker-voice information. *Bilingualism (Cambridge, England)*, *21*(3), 523–536. http://dx.doi.org/10.1017/S1366728917000153, URL https://pubmed.ncbi.nlm.nih.gov/29755282, Edition: 2017/06/09.

Levi, S. V., & Schwartz, R. G. (2013). The development of language-specific and language-independent talker processing. *Journal of Speech, Language, and Hearing Research*, *56*(3), 913–925. http://dx.doi.org/10.1044/1092-4388(2012/12-0095), URL https://pubs.asha.org/doi/abs/10.1044/1092-4388(2012/12-0095).

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*, 940–967a.

McGuire, G. L., & Babel, M. (2020). Attention to indexical information improves voice recall. In *INTERSpEECH* (pp. 1595–1599).

McLaughlin, D. E., Carter, Y. D., Cheng, C. C., & Perrachione, T. K. (2019). Hierarchical contributions of linguistic knowledge to talker identification: Phonological versus lexical familiarity. *Attention, Perception, & Psychophysics*, *81*(4), 1088–1107. http://dx.doi.org/10.3758/s13414-019-01778-5.

McLaughlin, D., Dougherty, S., Lember, R., & Perrachione, T. (2015). Episodic memory for words enhances the language familiarity effect in talker identification.

Neuhoff, J. G., Schott, S. A., Kropf, A. J., & Neuhoff, E. M. (2014). Familiarity, expertise, and change detection: Change deafness is worse in your native language. *Perception*, *43*(2–3), 219–222. http://dx.doi.org/10.1068/p7665.

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas - part II. *Language and Linguistics Compass*, *10*(11), 591–613.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. http://dx.doi.org/10.3758/BF03206860.

Orena, A. J., Polka, L., & Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience matters. *Journal of the Acoustical Society of America*, *145*(4), EL303–EL309. http://dx.doi.org/10.1121/1.5097735.

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. http://dx.doi.org/10.1016/j.cognition.2015.06.002, URL http://www.sciencedirect.com/science/article/pii/S0010027715300111.

Perrachione, T. K. (2018). Recognizing speakers across languages. In S. Frühholz, & P. Belin (Eds.), *The oxford handbook of voice perception*, Oxford University Press.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition*, *114*(1), 42–55. http://dx.doi.org/10.1016/j.cognition.2009.08.012, URL http://www.sciencedirect.com/science/article/pii/S0010027709002108.

Perrachione, T., Dougherty, S., McLaughlin, D., & Lember, R. (2015). The effects of speech perception and speech comprehension on talker identification. In *ICPhS*.

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *Journal of the Acoustical Society of America*, *146*(5), 3384–3399. http://dx.doi.org/10.1121/1.5126697.

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. http://dx.doi.org/10.1016/j.neuropsychologia.2006.11.015, URL http://www.sciencedirect.com/science/article/pii/S0028393206004611.

Senior, B., Hui, J., & Babel, M. (2018). Liu vs. Liu vs. Luke? Name influence on voice recall. *Applied Psycholinguistics*, *39*(6), 1117–1146. http://dx.doi.org/10.1017/S0142716418000267, URL https://www.cambridge.org/core/article/liu-vs-liu-vs-luke-name-influence-on-voice-recall/24B4DE90A2267926EF7A47A98B16CCDC.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. *Vol. 3*, In *Proceedings of the 17th international congress of phonetic sciences* (pp. 1846–1849). Hong Kong: URL https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/.

Stan Development Team (2021). Stan modeling language users guide and reference manual. URL http://mc-stan.org, R package version.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The "other-accent" effect in voice recognition. *Journal of Cognitive Psychology*, *24*(6), 647–653. http://dx.doi.org/10.1080/20445911.2012.675321.

Sumner, M., & Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encoding. *Journal of the Acoustical Society of America*, *134*(6), EL485–EL491.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics*, *77*, 1674–1684.

Theodore, R. M., & Flanagan, E. G. (2020). Determinants of voice recognition in monolingual and bilingual listeners. *Bilingualism: Language and Cognition*, *23*(1), 158–170.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131. http://dx.doi.org/10.1002/acp.2350010205, URL https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2350010205, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.2350010205.

Wester, M. (2012). 2017. *Speech Communication*, *54*(6), 781–790. http://dx.doi.org/10.1016/j.specom.2012.01.006, URL http://www.sciencedirect.com/science/article/pii/S0167639312000131.

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. http://dx.doi.org/10.1121/1.2913046.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *Journal of the Acoustical Society of America*, *137*(1), 419–432. http://dx.doi.org/10.1121/1.4904699.

Yu, M. E., Schertz, J., & Johnson, E. K. (2021). The other accent effect in talker recognition: Now you see it, now you don't. *Cognitive Science*, *45*(6), Article e12986.

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1), 11475. http://dx.doi.org/10.1038/srep11475.