Article in press – Authors' accepted manuscript
Please cite as:
Perrachione, T.K., Furbeck, K.T., & Thurston, E.J. (in press). Acoustic and linguistic factors affecting perceptual similarity judgments of voices. *Journal of the Acoustical Society of America.*

Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices

Tyler K. Perrachione*, Kristina T. Furbeck, Emily J. Thurston

Department of Speech, Language, & Hearing Sciences, Boston University, Boston, MA 02215

***Correspondence:**
Tyler K. Perrachione, Ph.D.
635 Commonwealth Ave.
Boston, MA 02215
Email: tkp@bu.edu

Abstract

The human voice is a complex acoustic signal that conveys talker identity via individual differences in numerous features, including vocal source acoustics, vocal tract resonances, and dynamic articulations during speech. It remains poorly understood how differences in these features contribute to perceptual dissimilarity of voices and, moreover, whether linguistic differences between listeners and talkers interact during perceptual judgments of voices. Here, native English- and Mandarin-speaking listeners rated the perceptual dissimilarity of voices speaking English or Mandarin from either forward or time-reversed speech. The language spoken by talkers, but not listeners, principally influenced perceptual judgments of voices. Perceptual dissimilarity judgments of voices were always highly correlated between listener groups and between forward/time-reversed speech. Using representational similarity analyses, we explored how acoustic features (fundamental frequency mean and variation, jitter, harmonics-to-noise ratio, speech rate, and formant dispersion) contributed to listeners' perceptual dissimilarity judgments, including how talker- and listener-language affected these relationships, finding the largest effects relating to voice pitch. Overall, these data suggest that, while linguistic factors may influence perceptual judgments of voices, the magnitude of such effects tends to be very small. Perceptual judgments of voices by listeners of different native language backgrounds tend to be more alike than different.

Keywords

voice perception; time reversal; perceptual dissimilarity; cross-language; speech perception; language-familiarity effect

PACS numbers

43.71.Bp Perception of voice and talker characteristics
43.71.Hw Cross-language perception of speech

I. INTRODUCTION

The human voice is a complex auditory stimulus that conveys a variety of information about a talker, most prominently who they are (their identity) and what they are saying (their linguistic message). Voices are a ubiquitous social and communicative signal, and a substantial literature has made forays into understanding how the various acoustic properties of the voice contribute to listeners' perceptual representation of a talker (e.g., Schweinberger & Zaske, 2018). However, little is known about whether the perceptual representation of voices, including the acoustic features that underlie talker identity, are affected by the language understood by listeners, the language spoken by talkers, or the interaction between them.

In the present work, we evaluate the hypothesis that listeners' perceptual space for voices is affected by their lifelong linguistic experiences (Fleming et al., 2014). From birth, listeners are inundated with experiences of voices speaking their native language and have extensive practice recognizing talkers of the same. However, listeners' experience with voices speaking in a foreign language is considerably less, and may be effectively nonexistent for certain foreign languages. Do different cultural experiences shape listeners' expertise with voice acoustics in such a way that they gain heightened perceptual sensitivity to subjective distinctions among the voices of individuals speaking in their native language compared to a foreign language? Effects of cultural experience have similarly been noted for perceptual sensitivity to own- vs. other-race faces (Meissner & Brigham, 2001), native vs. foreign phonetic contrasts (Werker & Tees, 1984), and pathological vs. healthy voice qualities (Kreiman, Gerratt, & Precoda, 1990), among many other domains of perceptual expertise. Ultimately, we are interested in whether experience-related differences in perceptual dissimilarity judgments of voices can help us discern the cognitive foundations of the language-familiarity effect in talker identification; that is, why listeners are more accurate at learning to identify talkers in their native language than in a foreign one (Goggin et al., 1991; Perrachione & Wong, 2007).

Previous research on voice processing has attempted to delimit the perceptual space for voices. Early work in this domain was based on highly subjective and qualitative judgments about vocal qualities, such as whether a talker sounds, for example, "harsh," "shrill," "monotonous," or "nasal" (reviewed in Kreiman, Vanlancker-Sidtis, and Gerratt, 2005). Contemporary clinical assessments of voice quality are born of this heritage, with voices rated on scales such as roughness, breathiness, strain, pitch, and loudness (Kempster et al., 2009). While the validity and reliability of clinical voice assessments are perhaps the most carefully scrutinized qualitative descriptors of the voice (e.g., Zraick et al., 2011; Karnell et al., 2007), they serve the specific purpose of helping clinicians identify the perceptual correlates of vocal pathology, rather than the perceptual correlates of individuals' vocal identity. More sophisticated efforts to identify the perceptual features that give rise to a talker's unique vocal identity come from studies looking for structure in listeners' perceptual similarity ratings (e.g., Baumann & Belin, 2010; Remez, Fellowes, & Nagel, 2007) and their relationship to voice acoustics. However, it does not strictly follow that, just because some acoustic dimensions are related to subjective judgments of voice dissimilarity, these same features need to be the ones that listeners use when recognizing a voice as familiar or when identifying a talker as a particular individual (Levi, 2018; Fecher & Johnson, 2018; Van Lancker & Kreiman, 1987; Perrachione, et al., 2014).

The question of perceived voice dissimilarity has recently been applied to studying the *language-familiarity effect in talker identification* (Fleming et al., 2014). In this extensively replicated phenomenon (reviewed in Perrachione, 2018), the ability to identify talkers by the sound of their voice is more accurate

when listening to one's native language than a foreign or less-familiar language. What makes the identity of native-language voices more memorable? A variety of factors have been suggested, including experience-specific prototypes for voices (Goggin et al., 1991), memories for voices abstracted from memories for speech (McLaughlin et al., 2015), and increased sensitivity to between-talker phonetic variation in one's native language (Perrachione et al., 2011). In their recent report, Fleming and colleagues (2014) suggested that the interaction between the languages spoken by listeners and talkers further extends to listeners' *dissimilarity judgments* of voices – i.e., their subjective, qualitative rating of how alike two voices sound – and, furthermore, that this language-familiarity effect in perceptual dissimilarity judgments was present even when voices had been time-reversed, rendering them incomprehensible. The authors of that study took this difference to mean that listeners are sufficiently sensitive to the phonological features of their native language, such that even in time-reversal, where the ability to identify wordforms is effectively eliminated, sufficient non-lexical but phonological information is preserved to facilitate native-language talker dissimilarity judgments.

However, this claim stipulates, but does not demonstrate, the persistence of language-specific phonological features in time-reversed speech. Whether such features persist or not is an unanswered empirical question, but there are many reasons to think that, if they do, they exist in a much more impoverished form than originally suggested. First, time-reversal does not preserve a language's phonological structure. When speech is time-reversed, the statistical relationships among phonemes and their order is demolished. For example, the time-reversed order of segments in the sentence, "A rod is used to catch pink salmon," (one of the Harvard Sentences (IEEE, 1969), which are regularly used in talker identification experiments) is [nmæs kŋɪp ʃtɛk ət dzuj zɪ dɑr ə], which contains numerous instances of segmental sequences that are phonologically unattested in English. Time-reversal of speech also destroys the temporal organization of subtle phonetic features, such as voice onset time, that contribute to the perception of voice identity (Ganugapati & Theodore, 2019). For instance, the time reversed version of "pink" in the sentence above will subject a listener to a physiologically impossible sequence of voicing: aspiration, burst, and then silence.

Thus, instead of the persistence of salient language-specific phonological or phonetic features in time-reversed speech, the observation of a language-familiarity effect for time-reversed talker dissimilarity ratings may instead implicate systematic differences between talkers of different languages in low-level acoustic factors that are sufficiently independent of speech content to be preserved in time-reversed stimuli. Listeners of the two languages therefore must have, through their lifelong experience with languages like either English or Mandarin, gained increased sensitivity to the relevant low-level acoustic features found in their native language, while losing sensitivity to the acoustic feature space of the other language. That is, listeners of one language putatively must not have the necessary experience-related perceptual sensitivity to access the distinguishing, non-linguistic, low-level acoustic differences in the other language.

It is not unreasonable to suspect that speakers of different languages will evince different low-level acoustic features broadly across their speech such that those differences will be preserved under time-reversal. English and Mandarin in particular are likely to differ on such basic acoustic dimensions as mean voice fundamental frequency and fundamental frequency variability, owing to the presence of syllable-level lexical tone contours in the latter but not the former (Shih, 1988). Similarly, Mandarin and English differ in their prosodic organization, such that Mandarin is a syllable-timed language whereas English is stress-timed (e.g., Mok, 2008), leading to differences in not only the duration of syllables but also their relative amplitude, both of which are low-level, non-segmental features that would be preserved in time-

reversal. Listeners of different language backgrounds may also be differentially sensitive to acoustic correlates of voice quality (Keating & Esposito, 2007; Kreiman, Gerratt, & ud Dowla Khan, 2010); Mandarin and English make differential use of creaky voice to signal the third (dipping) tone in Mandarin (Davison, 1991) and to signal either utterance finality or an allophone of /t/ in English (Slifka, 2007). Thus, if language-specific phonological features are not the source of listeners' biases in perceiving talker differences in time-reversed speech, perhaps it is possible instead to attribute these language-based differences to different sensitivity to low-level acoustic features.

Finally, the results of the Fleming and colleagues (2014) report also imply two additional hypotheses: that the language-familiarity effect in perceptual dissimilarity judgments should be *stronger* for time-forward voices (where additional language-specific acoustic and phonetic details are present, thus giving listeners a stronger perceived difference in their native language) and that listeners of different language backgrounds should rely on different low-level acoustic features when making perceptual dissimilarity judgments of voices.

Our aims for this report are therefore to replicate and extend the results of Fleming and colleagues (2014) in several ways. First, we attempted a veridical replication of the prior report by repeating their methods and statistical analyses as closely as possible, while using new stimuli and new participants. Second, we wanted to explore how this perceptual dissimilarity space differed between *time-reversed* voices (a fairly unnatural stimulus) and *time-forward* ones. Third, we wanted to understand whether listeners' native language affected their dissimilarity judgments when two recordings came from the *same* speaker (cf. Lavan et al., 2018; 2019a; 2019b), not just when they came from *different* speakers. Fourth, we aimed to look at not just whether listeners' native language backgrounds led to *differences* in their perception of talker dissimilarity, but also whether there were any fundamental *similarities* in the perceptual space for talkers among listeners of different language backgrounds. Finally, we wanted to explore the acoustic-phonetic factors that may affect perception of voice dissimilarity, and see how these factors differ (1) across talkers' languages; (2) across listeners' languages; (3) in forward vs. time reversal, and (4) any interaction between these levels.

In this study, we asked native English- and Mandarin-speaking listeners to rate the perceptual dissimilarity of pairs of voices speaking either English or Mandarin, that were played either forward or time-reversed. We also made acoustic measurements on these voices for both vocal source features (fundamental frequency (f_0) mean and variation, and voice quality (jitter and harmonics-to-noise ratio (HNR)) and vocal filter / articulatory features (formant dispersion and speech rate). The choice of these features was motivated by the likelihood of their being preserved between forward and time-reversed speech, as well as their potential to be attested – or attended to – differently between Mandarin and English. We then examined both the divergence and convergence of listeners' perceptual dissimilarity judgments across differences in talkers' and listeners' linguistic background. We further assessed the relationship between listeners' perceptual dissimilarity judgments and the acoustic features of the voice samples, including how these were affected by talker language, listener language, and their interaction.

Ultimately, the results of this study reveal that listeners' perceptual dissimilarity judgments of voices are highly similar regardless of the native language of the listeners, the native language of the talkers, or whether the voices are played time-forward or time-reversed. Furthermore, the primary acoustic feature associated with perceptual dissimilarity judgments is the highly salient, reversal-invariant, language-independent mean fundamental frequency of a talker's voice. These results reveal that the perceptual dissimilarity space for voices tends to be conserved across listeners of different language backgrounds,

even for highly disruptive manipulations of the voice stimuli, such as what language they speak or whether they are time reversed. This outcome calls into question the potential contribution that studying perceptual dissimilarity judgments of voices can make toward revealing the cognitive foundations of the language-familiarity effect in talker identification. Thus, we conclude with a discussion of the strengths and weaknesses of perceptual dissimilarity judgments as a paradigm for studying voice cognition, and we propose a framework for future research in this domain.

II. METHODS

In this study, two groups of listeners (native speakers of American English or Mandarin Chinese) listened to pairs of recordings of speech in English and/or Mandarin and rated the perceptual dissimilarity of the voices in the two recordings. Listeners completed this task under either of two conditions: *time-reversed speech* (cf. Fleming et al., 2014), in which the recordings were played backwards and were incomprehensible, and *forward speech*, in which natural speech was heard. Acoustic properties of talkers' speech were also measured. Listeners' perceptual dissimilarity judgments were analyzed for effects of listener- and talker-language, and time-reversal, as well as their relationship to acoustic features.

A. Participants

Participants in this study included native speakers of American English ($N = 40$; 35 female, 5 male; ages 18-24, mean = 20.3 years) and native speakers of Mandarin Chinese ($N = 40$; 32 female, 8 male; ages 18-37, mean = 21.2 years). The native English speakers had no familiarity with Mandarin; the native Mandarin speakers, who were born and raised in China but currently living or studying in the United States, were bilingual in English. Mandarin participants reported exposure to English beginning on average at age 6 (±2.5 years, range 1-12 years old) and having on average 13.3 ± 4.3 years of English-language study (range 3-30 years). Of the Mandarin participants, 33 reported currently using Mandarin more than English, despite living in the United States, and 3 reported using the two languages equally. All participants indicated a history free from speech, language, or hearing disorders. All participants provided informed, written consent prior to undertaking the experiment. This study was approved and overseen by the Institutional Review Board at Boston University. Of the 40 participants in each group, 20 completed the task in the *time-reversed speech* condition and 20 completed the task in the *forward speech* condition. The size of both the participant and item samples in each condition were identical to those of Fleming and colleagues (2014).

B. Stimuli

A total of 400 recordings from 40 female speakers were presented to listeners in this study. We recorded female native speakers of American English ($N = 20$; ages 18-29, mean = 23.1 years) reading List 2 of the phonetically balanced English "Harvard Sentences" (IEEE, 1969) and 20 female native speakers of Mandarin (N = 20; ages 18-30, mean = 23.2 years) reading sentences 1-4 and 6-11 of the Mandarin Speech Perception Test (Fu et al., 2011). Speakers were recorded in a sound attenuated booth using a Shure MX153 earset microphone, a Behringer Ultragain Pro MIC2200 2-channel tube microphone preamplifier, and Roland Quad Capture USB audio interface with a sampling rate of 44.1 kHz and 16-bit digitization.

The recording of each sentence was cut to 1250 ms from its onset (following Fleming et al., 2014), with a linear amplitude ramp applied to the final 125 ms to avoid the abrupt, unnatural sound of a cut

recording. Cut recordings were RMS amplitude normalized to 68 dB SPL. In the *time-reversed speech* condition, recordings of stimuli were presented backwards; in the *forward speech* condition, the natural speech recordings were presented. Stimulus editing was completed in Praat.

C. Procedure

The study took place in a sound attenuated booth. Stimulus presentation was controlled using PsychoPy v1.83.03 (Pierce, 2007). Recordings were presented via Sennheiser HD 380 Pro circumaural headphones. On each trial, the listener heard a pair of recordings and was asked to rate how similar or different the voices sounded on an analog sliding scale ranging from 0 (indicating absolute certainty that the voices were the same) to 1 (indicating absolute certainty that the voices were different). Listeners were encouraged to use the full extent of the scale and not just the endpoints; they were told that we were studying the voices themselves and not listeners' accuracy at telling them apart. Participants were given the option to replay the pair of recordings on each trial as many times as they needed before submitting their response. Participants had no prior exposure to the stimuli or voices used in this study. The experimental procedure is schematized in **Figure 1**.

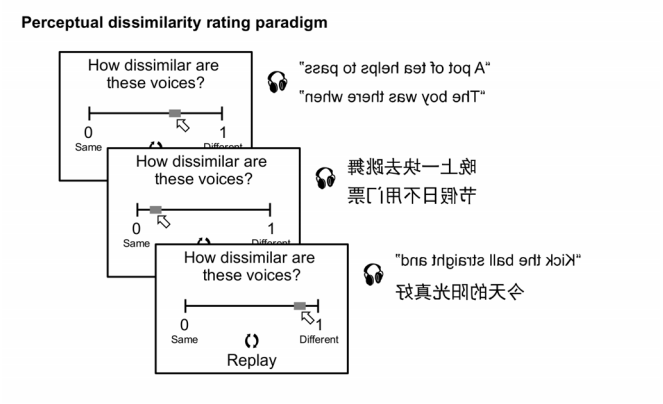

Figure 1: Perceptual dissimilarity rating paradigm. On each trial, participants heard recordings of two different speech samples and indicated how dissimilar the voices in the two recordings sounded by sliding a selector along an ordinal scale from 0 (definitely the same voice) to 1 (definitely different voices). Participants were encouraged to use the entire scale. Recordings came from either the same talker in one language, different talkers in the same language, or different talkers in different languages. For half of the listeners, the recordings were time-reversed, as in Fleming et al. (2014) (indicated here by mirror-reversal of the text); for the other half, the recordings were presented naturally.

Listeners heard all possible combination of talkers, resulting in a total of 820 trials (40 *same-identity* trials, in which the two recordings were spoken by the same talker; 190 *native-language* trials, in which pairs of recordings from all combinations of the 20 talkers in the listener's native language (English or Mandarin) were presented; 190 *foreign-language* trials, in which pairs of recordings from all possible combinations of the 20 talkers in the listener's non-native language were presented; and 400 *cross-language* pairs, in which pairs of recordings from all combinations of talkers from the two languages were presented). For all pairs of recordings, the sentences spoken by the two talkers were different. All conditions were presented during each session, and the order of trials was randomized. Each participant heard a unique set of talker-sentence pairs, as well as a unique order of trials. See "Open Source Dataset" for the full experimental materials (recordings, stimulus pairs, experiment scripts) available online.

The study was self-paced and took approximately 2 hours to complete. The program was broken into 4 sessions, consisting of 205 trials each. Listeners were allowed to take a break between each session. Participants were allowed to complete the study across two consecutive days, completing two sessions during each visit. Twenty-one participants completed the task in one visit; 59 participants completed the study in two visits.

In conducting this study, we aimed to replicate the design of Fleming and colleagues (2014) exactly. To the best of our ability, we have done so with the following exceptions or additions: (i) Additional groups

of Mandarin and English listeners also rated the dissimilarity of talkers from forward speech, whereas Fleming and colleagues used only time-reversed speech; (ii) listeners in our study had no prior exposure to the voices, whereas listeners in the 2014 study had been exposed to those voice stimuli during a prior experiment of unspecified design; (iii) we investigated listeners' dissimilarity biases from trials in which the same talker was heard twice, not just trials in which different talkers were heard; (iv) we performed additional analyses looking into the similarities, not just differences, between listeners' perceptual judgments across native language backgrounds; and, (v) we performed representational similarity analyses to ascertain whether listeners' perceptual dissimilarity judgments were related to a variety of acoustic features, and whether these differed with respect to talker-listener language pairings or time-reversal.

D. Acoustic measurements

To investigate the relationship between listeners' perceptual dissimilarity judgments and the acoustic properties of talkers' speech, we analyzed a number of acoustic features that can reflect differences in vocal source acoustics, vocal filter acoustics, and speech articulation. These features were selected based on their previous implication as potentially perceptually distinguishing acoustic features of voices (e.g., Baumann & Belin, 2008; Latinus & Belin, 2011b; Latinus et al., 2013; Remez, Fellowes, & Nagel, 2007; Schweinberger et al., 2014), because they may have differential attestation based on talker language – or differential attention based on listener language – between Mandarin and English, and because they are likely preserved between forward and time-reversed speech. A number of these features (mean f_0, f_0 range, formant dispersion, and HNR) were also reported for the speakers in Fleming and colleagues' (2014) stimuli; however, the relationship between speech acoustics and perceptual dissimilarity ratings was not examined in that report. All acoustic measurements were accomplished in Praat.

Fundamental frequency (f_0). We measured the mean and standard deviation of talkers' voice fundamental frequency from each stimulus recording. The standard pitch tracking parameters in Praat were used, unless these resulted in pitch tracking errors, such as pitch doubling or pitch halving, owing to misidentification of the relevant waveform peak by the autocorrelation function. The pitch contour of every recording was visually inspected overlaid on the spectrogram to identify any such errors, in which case the minimum and maximum pitch range for that particular recording was adjusted to eliminate any such error.

Voice quality. Two measures of voice quality were obtained for each recording: *jitter* and *harmonics-to-noise ratio (HNR)*. Jitter is an acoustic correlate of temporal perturbation in vocal fold vibration and is perceptually related to the creakiness of a voice (Karnell et al., 2007). We used the five-point period perturbation quotient algorithm to estimate jitter in each recording, as this algorithm provides an estimate of vocal temporal perturbation that is robust to ongoing pitch dynamics in natural speech (Davis, 1981). Jitter is expressed as a mean percent difference in cycle-to-cycle periodicity. HNR is an acoustic correlate of voice quality that reflects relative energy of the periodic and aperiodic components of the voice, expressed in dB. Briefly, both jitter and HNR provide indices of the extent to which a talker's voice quality differs from the modal voice. Voice quality measurements were made simultaneously with, and using the same pitch settings as, the fundamental frequency measurements.

Speech rate: As each recording was truncated at 1.25 seconds, sometimes mid-syllable, we determined talkers' speech rate as the number of full syllables listeners heard in each recording, divided by the time it took to produce those syllables. Counting up until the time of the end of the last full syllable in each recording, we calculated talkers' speech rate for each utterance in syllables per second.

Formant dispersion: Formant dispersion is a measure of vocal tract length, with shorter vocal tracts producing higher frequency resonances and thus greater frequency distance between formants (Fitch, 1997). Formants were measured using the standard settings in Praat, adjusted as needed on a recording-by-recording basis to obtain good formant tracking. Values for each of the first four formants (F1-F4) were extracted across the entire recording. Because Praat is prone to formant tracking errors at the transition between the open and closed vocal tract at syllable boundaries in natural, running speech, we weighted the contribution of each sample by the degree of vocal tract opening at that time point. We calculated the mean frequency of each formant across the utterance weighted by the intensity contour, such that formant values measured during the maximally open vocal tract (i.e., high-intensity speech) contributed more to the average than those measured proximal to closures (i.e., low-intensity speech). This method allowed us to make use of the entire utterance, as heard by listeners during the voice judgment task, while simultaneously maximizing the signal-to-noise ratio of formant measurements by reducing the amount of measurement error due to poor formant tracking during vocal tract closures. Formant dispersion was calculated using the mean of the differences of adjacent formants (Fitch, 1997).

E. Statistical analyses

Data were analyzed in R v3.5.1 using the packages *ez* for repeated-measures analyses of variance (ANOVA) and *lme4* and *lmerTest* for linear mixed effects models. The fixed and random effects structure of each model is described below. The significance of factors in the linear mixed models was determined using Type-III ANOVAs incorporating Satterthwaite's method for approximating denominator degrees of freedom. Post-hoc comparisons to identify the direction and source of main and interaction effects from the ANOVA were conducted using difference of least-square means implemented via the function *difflsmeans*. We adopted a significance criterion of $\alpha = 0.05$ for ANOVAs and other planned comparisons, and applied Bonferroni-corrected alpha criteria when assessing significance of post-hoc tests on each model.

III. RESULTS

We first attempted to replicate the observation of a language-familiarity effect for talker dissimilarity ratings from time-reversed speech reported by Fleming and colleagues (2014). Using only the data from listeners who made judgments of time-reversed recordings, we performed the same statistical analyses described in the prior report. Second, we analyzed the full dataset we collected, including all 65,600 dissimilarity judgments performed by English- and Mandarin-speaking listeners on all voice pairs from both time-reversed and forward speech, for differences related to talker- or listener-specific factors. Third, we examined whether listeners' perceptual dissimilarity judgments across the entire dataset were related to acoustic differences between the pairs of recordings and, if so, whether these relationships differed across talker language, listener language, or time-reversal.

A. Attempted replication of a language-familiarity effect for dissimilarity judgments of time-reversed recordings

First, a repeated-measures analysis of variance (ANOVA) was conducted on the dependent measure of listeners' dissimilarity ratings, with the within-subject factor of *pair type* (same-talker pairs, foreign-language pairs, native-language pairs, and cross-language pairs). This analysis revealed a significant effect of *pair type* ($F(3,117) = 346.95$, $p \ll 0.0001$, $\eta^2_G = 0.789$), paralleling the prior report (**Figure 2A**). Same-talker pairs were rated the least dissimilar (larger numbers indicate greater mean dissimilarity; mean ±

across-participants standard error: 0.13 ± 0.02), then foreign-language pairs (0.63 ± 0.02), native-language pairs 0.65 ± 0.02) and finally cross-language pairs were rated most dissimilar (0.78 ± 0.02).

Post-hoc tests revealed that cross-language pairs were rated as significantly more dissimilar than native-language pairs, foreign language pairs, and same-talker pairs (all paired $t(39) > 8.95$, all $p \ll 0.0001$). However, in contrast to the prior report, native-language pairs were not rated as significantly more dissimilar than foreign-language pairs ($t(39) = 0.91$, $p = 0.37$), though they were more dissimilar than same-talker pairs ($t(39) = 18.68$, $p \ll 0.0001$). Foreign-language pairs were also rated as significantly more dissimilar than same-talker pairs ($t(39) = 18.68$, $p \ll 0.0001$).

Second, we analyzed the native- and foreign-language talker pairs separately for native speakers of English (native pairs: 0.62 ± 0.03; foreign pairs: 0.68 ± 0.02) and Mandarin (native pairs: 0.68 ± 0.04; foreign pairs: 0.59 ± 0.04) (**Figure 2B**). These data were submitted to a 2x2 repeated-measures ANOVA, with *listener language* (English, Mandarin) as the between-subjects factor and *talker language* (English, Mandarin) as the within-subjects factor. Like Fleming and colleagues (2014), we found no main effect of *listener language* ($F(1,38) = 0.073$, $p = 0.79$, $\eta^2_G = 0.0018$). However, unlike the prior report, we did not find a *listener language* × *talker language* interaction ($F(1,38) = 2.44$, $p = 0.13$, $\eta^2_G = 0.0025$), but we did find a main effect of *talker language* ($F(1,38) = 75.91$, $p \ll 0.0001$, $\eta^2_G = 0.071$). Post-hoc tests revealed that the Mandarin talkers were perceived as more dissimilar than the English talkers by both the Mandarin listeners (paired $t(19) = 6.31$, $p \ll 0.0001$) and by the English listeners (paired $t(19) = 6.16$, $p \ll 0.0001$).

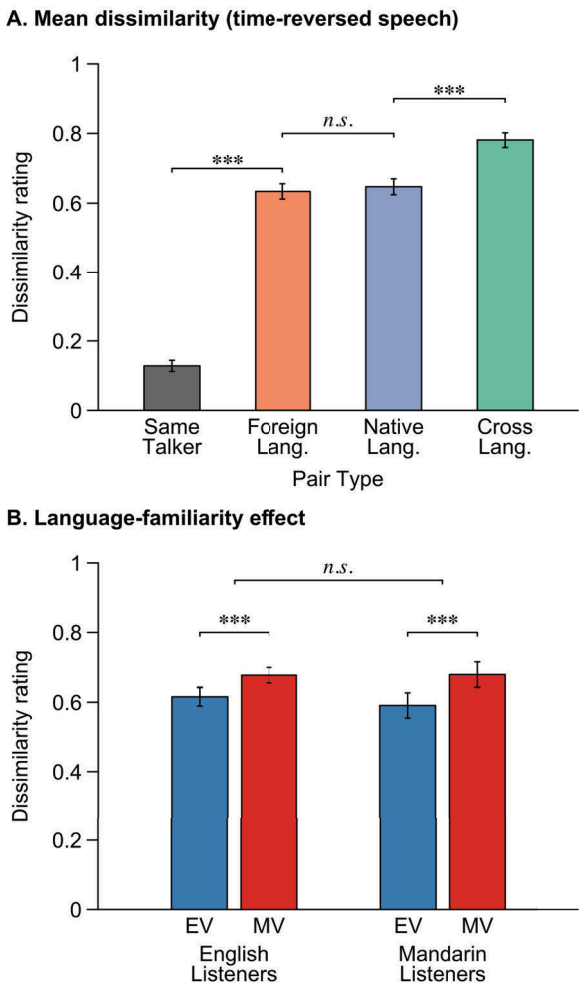

Figure 2: Perceptual dissimilarity of time-reversed talker pairs. Panel **(A)** shows the mean dissimilarity rating for each talker pair type across both listener groups for time-reversed recordings, following the conventions of Fleming et al. (2014). Listeners rated same talker pairs as least dissimilar and cross-language talker pairs as most dissimilar. Different-talker pairs speaking the same language were rated similarly, regardless of whether they were speaking listeners' native or foreign language. Panel **(B)** shows dissimilarity ratings of English (EV) and Mandarin (MV) talker pairs separately for native English- and Mandarin-speaking participants. Both groups found Mandarin talker pairs more dissimilar, and there was no talker language × listener language interaction, suggesting the language-familiarity effect does not influence talker dissimilarity ratings from time-reversed speech.

B. Dissimilarity judgments of forward and time-reversed talkers

We next analyzed the full dataset, including all 65,600 dissimilarity judgments participants made (including for all time-reversed and forward recordings, all talker pairs, and by listeners of both native languages). The pattern of mean dissimilarity judgments across participants for each pair of talkers is shown in **Figure 3A** (for time-reversed recordings) and **Figure 3B** (for forward recordings). Each cell of a dissimilarity matrix corresponds to a unique pair of talkers, and the mean dissimilarity rating for that pair is indicated by the amount of shading, from 0 (most similar, darkest) to 1 (most dissimilar, lightest); the quadrants of the matrices correspond to English-speaking voice pairs (top left), Mandarin-speaking voice pairs (bottom-right), and cross-language pairs (bottom left and top right). Same-talker pairs are depicted along the diagonal.

Perceptual dissimilarity judgments are inherently non-Gaussian, given the distribution is bounded by 0 and 1. Inspection of the distribution of responses further revealed substantial deviation from normality (**Figure 3C**), with responses clustered near 0 and 1 (Anderson-Darling test of normality, $A = 6395.4$, $p \ll 0.0001$). Correspondingly, we applied arcsine transformation (Studebaker, 1985) to the dissimilarity rating data prior to inferential statistics, after which the data were not as skewed towards the extrema of the scale, but were nonetheless still not normally distributed ($A = 4094.5$, $p \ll 0.0001$). Listeners exhibited a strong preference for dissimilarity rankings at the extrema of the range: Across all the different-talker trials (foreign, native, and cross-language) in our data, fully 50% of responses had dissimilarity ratings of ≥ 0.93 (where 1 meant "definitely different"), and nearly 42% of responses were "1" exactly. For same-talker trials, 56% received a dissimilarity rating of "0" exactly (where 0 meant "definitely the same"), and only 23% of same-talker trials had a rating > 0.1. Only 17% of all trials were rated within the middle half of the range (0.25-0.75).

Participants' arcsine-transformed dissimilarity ratings were submitted to a linear mixed effects model with fixed factors including *pair type* (same-talker pairs, foreign language different-talker pairs, native language different-talker pairs, and cross-language pairs), *listener native language* (English, Mandarin) and *recording direction* (time-reversed, forward). Random factors in the model included by-participant intercepts and by-participant slopes for the within-subject effect of *pair type*, as well as random item intercepts for each (unordered) pair of talkers.

An ANOVA of this model revealed a significant main effect of *pair type*, and significant *pair type* × *listener native language* and *pair type* × *recording direction* two-way interactions **(Table 1)**. The other main, two-, and three-way interaction effect terms were not significant. Post-hoc pairwise tests revealed that same-talker pairs were rated as less dissimilar than foreign, native, or cross-language pairs (all $t > 19.64$, $p \ll 0.0001$). Likewise, cross-language pairs were rated as more dissimilar than native or foreign pairs (both $t > 6.96$, $p \ll 0.0001$). Native-language pairs were also rated as significantly more dissimilar than foreign-language pairs ($t = -5.28$, $p \ll 0.0001$). However, given that two other factors had significant interactions with *pair type*, these simple effects require further elaboration.

The *pair type* × *recording direction* interaction was driven by participants' tendency to give higher dissimilarity ratings for native-language talkers when hearing forward vs. time-reversed recordings ($t = -3.29$, $p = 0.0015$), but less so for cross-language pairs ($t = -1.53$, $p = 0.13$), foreign-language pairs ($t = -1.83$, $p = 0.071$), or same-talker pairs recordings ($t = 1.43$, $p = 0.16$) – the latter of which had the opposite tendency, with even lower dissimilarity ratings from forward speech (**Fig. 3D**).

The *pair type* × *listener native language* interaction was driven by a tendency for higher dissimilarity ratings on native-language pairs by native Mandarin listeners compared to native English listeners

(t = -2.98, p = 0.0037), but not for same-talker pairs (t = 0.36, p = 0.72), foreign-language pairs (t = 0.63, p = 0.53), or cross-language pairs (t = 0.45, p = 0.65). Mandarin listeners tended to rate native-language pairs as more dissimilar than foreign-language pairs (t = -6.76, $p \ll 0.0001$), whereas English listeners' ratings tended to be in the opposite direction (t = 1.72, p = 0.086). The two listener groups did not differ overall in their perceived dissimilarity of English talker pairs (English native pairs vs. Mandarin foreign pairs; t = -0.11, p = 0.91), but did differ in their perceived dissimilarity of Mandarin talker pairs (t = -2.38, p = 0.02), such that Mandarin listeners judged these talker pairs as more dissimilar than English listeners did.

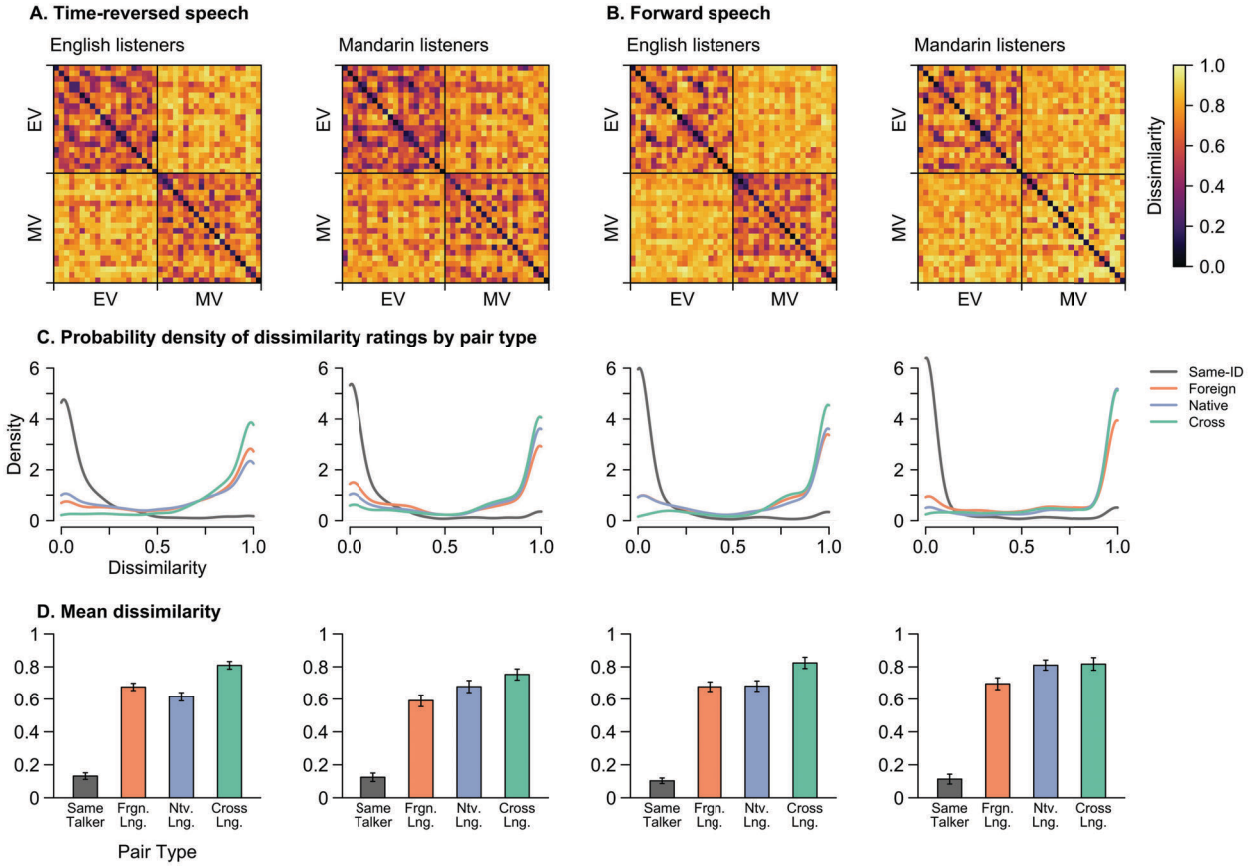

Figure 3: Patterns of perceptual dissimilarity judgments for all talker pairs across time-reversed or forward speech and listeners' native language. In the matrices at top, each row and column correspond to an individual talker, such that each cell indicates the perceived dissimilarity between each pair of talkers from 0 (maximally similar) to 1 (maximally dissimilar). Same-identity talker pairs occur along the diagonal; the top left quadrant contains pairs of English-speaking talkers (EV); the bottom right quadrant contains pairs of Mandarin-speaking talkers (MV); and the top right and bottom left quadrants contain cross-language talker pairs. Panels in **(A)** show the pattern of dissimilar ratings across all time-reversed talker pairs for native English (left) and Mandarin (right) listeners. Panels in **(B)** show the corresponding pattern of dissimilar ratings across all time-forward talker pairs. The color scale indicates the mean dissimilarity across listeners for a talker pair, with darker colors corresponding to less dissimilarity and lighter colors corresponding to greater dissimilarity. Panels in **(C)** show the probability density functions of participants' dissimilarity ratings for each pair type corresponding to the matrix directly above. The area under each curve is 1. Panels in **(D)** show the mean dissimilarity rating for each condition corresponding to the matrix and density plot above. Error bars are ± s.e.m. across participants. Note the low dissimilarity rating along the diagonal of each matrix for the same-talker pairs, reflected in the leftmost peak (grey lines) in the density plots and the low mean dissimilarity in the barplots. Note also the relatively higher dissimilarity rating for talker pairs in the cross-language quadrants (top right, bottom left), and the corresponding rightmost peak of the cross-language (green) distribution below. Reading across the matrices, note the strikingly similar pattern of cell-level similarity ratings between the two listener groups and between time-reversed and forward speech. These similarities are considered quantitatively in **Fig. 5**.

Table 1. Linguistic factors affecting perceptual dissimilarity judgments of voices (all data).

Fixed factor	***F***	***df*** **(n,d)**	***p*****-value**
Pair type	172.41	(3, 121)	$\ll 0.0001$
Listener native language	0.34	(1, 76)	0.56
Recording direction	3.73	(1, 76)	0.057
Pair type × Listener native language	10.71	(3, 114)	$\ll 0.0001$
Pair type × Recording direction	6.93	(3, 76)	0.00035
Listener native language × Recording direction	1.26	(1, 76)	0.27
Pair type × Listener native language × Recording direction	0.86	(3, 76)	0.46

1. The language-familiarity effect

Organizing the data by pair type may obscure some potentially interesting relationships between talker language, listener language, and talker identity. In particular, treating languages as "native" or "foreign" may miss main effects due to language (English vs. Mandarin), or listener by talker language interactions. Similarly, by treating all same-talker pairs as a single category, we forego the ability to detect effects of talkers' or listeners' language on how listeners "tell voices together" (Lavan et al., 2019a; 2019b) compared to telling them apart. We therefore performed two additional planned analyses on listeners' dissimilarity ratings, the first from only the native- and foreign-language pairs, now organized by English and Mandarin talkers, and the second for only the same-talker pairs, likewise organized by language.

a. Different-talker pairs

Participants' arcsine-transformed dissimilarity ratings for pairs of different English- or Mandarin-speaking talkers were submitted to a linear mixed effects model with fixed factors including *talker language* (English, Mandarin), *listener native language* (English, Mandarin) and *recording direction* (time-reversed, forward). Random factors in the model included by-participant intercepts and by-participant slopes for the within-subject effect of *talker language*, as well as random item intercepts for each pair of talkers.

An ANOVA of this model revealed significant main effects of *talker language* and of *recording direction* and a significant *talker language* × *listener native language* interaction **(Table 2)**. There was also a significant three-way *talker language* × *listener native language* × *recording direction* interaction. The other main effects and interactions were not significant.

The main effect of *talker language* was driven by overall higher dissimilarity ratings for Mandarin talker pairs ($t = 3.95$, $p \ll 0.0001$). The main effect of *recording direction* was driven by overall higher dissimilarity ratings for forward vs. time-reversed recordings ($t = 2.63$, $p = 0.010$). The *talker language* × *listener native language* interaction represents the same differences giving rise to the *pair type* × *listener native language* interaction in the previous model – namely, English- and Mandarin-speaking listeners differ in their judgments of Mandarin-speaking voices, but not English-speaking ones.

Exploring the three-way interaction reveals the lack of consistent attestation of a language-familiarity effect in talker dissimilarity ratings **(Figure 4A)**. If language familiarity affects listeners' perceptual dissimilarity space for talkers, then native-language talker pairs should reliably be rated as more dissimilar than foreign-language talker pairs. However, post-hoc pairwise tests revealed that this was not always the case: Contrary to the language-familiarity hypothesis, English listeners actually rated *Mandarin*-speaking

talkers as more dissimilar than English-talker pairs when listening to time-reversed speech (t = 2.94, p = 0.0035), and they did not differ in their ratings of talkers of the two languages from time-forward speech (t = -0.41, p = 0.68). Mandarin listeners, however, did rate Mandarin-talker pairs as more dissimilar than English-talker pairs from both time-reversed (t = 4.23, $p \ll 0.0001$) and forward recordings (t = 5.75, $p \ll$ 0.0001).

Table 2. Linguistic factors affecting perceptual dissimilarity judgments of different-talker pairs.

Fixed factor	***F***	***df* (n,d)**	***p*-value**
Talker language	15.61	(1, 414)	≪ 0.0001
Listener native language	1.63	(1, 76)	0.21
Recording direction	6.89	(1, 76)	0.010
Talker language × Listener native language	27.88	(1, 76)	≪ 0.0001
Talker language × Recording direction	1.68	(1, 76)	0.20
Listener native language × Recording direction	1.41	(1, 76)	0.24
Talker language × Listener native language × Recording direction	11.97	(1, 76)	0.0009

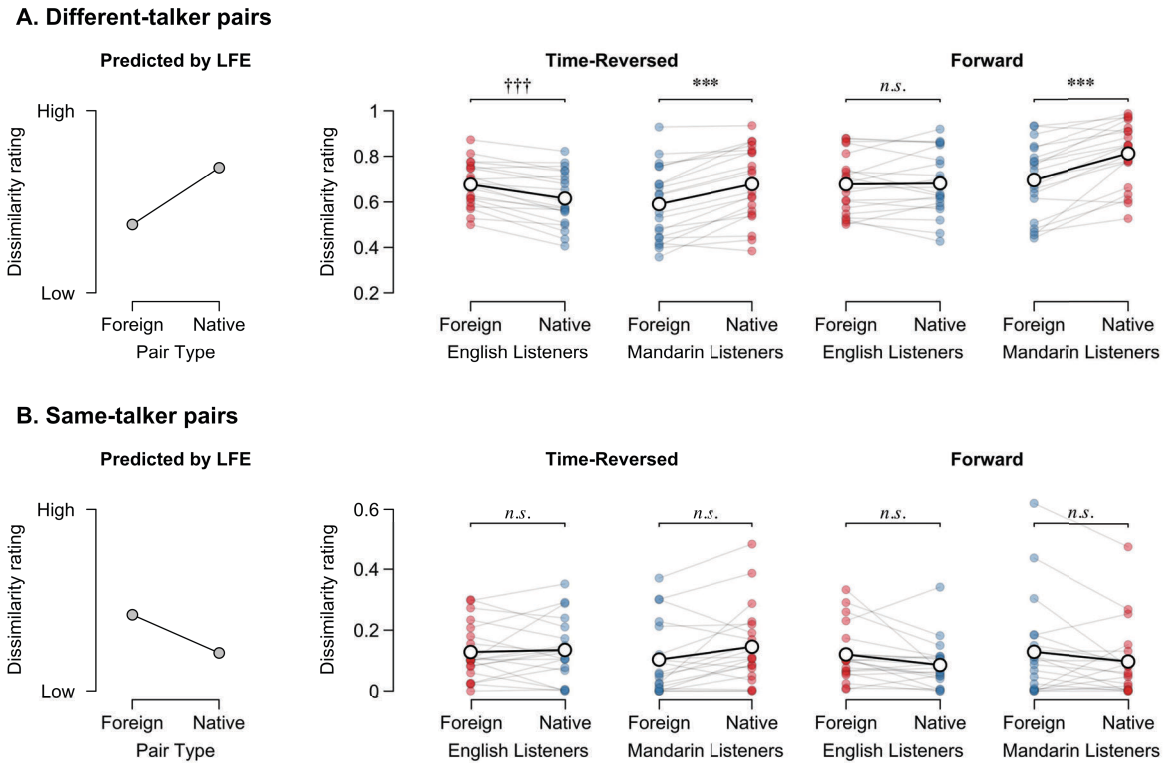

Figure 4: Patterns of between-language dissimilarity rating differences are infrequently consistent with the predictions of the language-familiarity effect. (A) Predicted (left) and measured (right) dissimilarity ratings for different-talker pairs in each language, recording direction, and listener group. **(B)** Predicted (left) and measured (right) dissimilarity ratings for same-talker pairs in each language, recording direction, and listener group. *Legend:* Points represent mean dissimilarity ratings of individual participants in each condition; lines connect points from the same participant to show the direction of the effect. Red points indicate Mandarin talkers, blue points indicate English talkers; points are partially transparent to reveal overlap. Larger white points show the mean dissimilarity rating across participants in each condition. *Symbols & abbreviations:* LFE, language-familiarity effect; *n.s.* $p > 0.0125$; *** $p < 0.005$ in the direction predicted by the language-familiarity effect; ††† $p < 0.005$ in the opposite direction of the predictions of the language-familiarity effect.

b. Same-talker pairs

Participants' arcsine-transformed dissimilarity ratings for pairs of recordings from the same English or Mandarin talkers were submitted to a linear mixed effects model with the same structure as that for different-talker pairs. An ANOVA of this model revealed only a significant three-way *talker language* × *listener native language* × *recording direction* interaction **(Table 3)**. All the other main and interaction effects were not significant.

A language-familiarity effect on listeners' perceptual dissimilarity judgments for same-talker pairs make the opposite prediction of their judgments for different-talker pairs: namely, listeners should be more sensitive to the fact that two recordings come from the same talker in their native language, thus providing *lower* dissimilarity ratings for same-talker pairs in their native language than in a foreign language. Exploring the three-way interaction provides little evidence for an effect of language familiarity on judging the same talker to sound more similar to herself **(Figure 4B)**. Post-hoc pairwise tests revealed that, contrary to the language-familiarity hypothesis, Mandarin listeners actually tended to rate pairs of recordings from a single Mandarin-speaking talker as more dissimilar than pairs of recordings from a single English-speaking talker when listening to time-reversed speech (t = 1.81, p = 0.074), but did tend to rate same-talker pairs in the expected direction for time-forward speech (t = -1.76, p = 0.082), though not significantly so in either case. English listeners also did not exhibit a pattern of dissimilarity ratings predicted by language familiarity: they did not rate English and Mandarin same-talker pairs differently either for time-reversed recordings (t = -0.29, p = 0.78) or forward speech (t = 2.24, p = 0.028 ($\alpha_{Bonf.}$ = 0.0125)).

Table 3. Linguistic factors affecting perceptual dissimilarity judgments of same-talker pairs.

Fixed factor	***F***	***df* (n,d)**	***p*-value**
Talker language	0.59	(1, 40)	0.45
Listener native language	0.13	(1, 76)	0.72
Recording direction	2.05	(1, 76)	0.16
Talker language × Listener native language	1.18	(1, 76)	0.28
Talker language × Recording direction	0.35	(1, 76)	0.55
Listener native language × Recording direction	0.09	(1, 76)	0.76
Talker language × Listener native language × Recording direction	12.14	(1, 76)	0.0008

C. Consistent perceptual dissimilarity judgments across listener groups and time-reversal

Despite the subtle differences in magnitude noted above, the overall patterns of perceptual dissimilarity judgments for native- and foreign-language talkers tended to be highly similar across listener groups and time-reversal. Bivariate correlations of the mean perceptual dissimilarity across listeners of each different-talker pair revealed that the pattern of perceptual dissimilarity for time-reversed English voices (i.e., the lower triangle of the English-English quadrants in **Fig. 3A**; shown in **Fig. 5A**) was significantly correlated between English and Mandarin listeners (r_{188} = 0.85, $p \ll 0.0001$). Likewise, the corresponding pattern of perceptual judgments for the time-reversed Mandarin voices was also highly correlated across listener groups (r_{188} = 0.78, $p \ll 0.0001$). For the forward voices, the two listener groups likewise exhibited extremely similar patterns of dissimilarity judgment for pairs of English (r_{188} = 0.80, $p \ll 0.0001$) and Mandarin (r_{188} = 0.65, $p \ll 0.0001$) talkers.

The pattern of listeners' dissimilarity judgments was also significantly related across the time-reversal manipulation. Native English-speaking listeners tended to find the same English-speaking talkers

to be more or less dissimilar regardless of whether their speech was comprehensible or not (i.e., the lower triangles of the English-English quadrants in **Fig. 3A and 3B**; shown in **Fig. 5B**) ($r_{188} = 0.79$, $p \ll 0.0001$). The pattern of dissimilarity judgments elicited from English listeners was also highly correlated for Mandarin voices, regardless of time reversal ($r_{188} = 0.75$, $p \ll 0.0001$). For native Mandarin-speaking listeners, as well, the same English-speaking voices were more or less dissimilar regardless of time reversal ($r_{188} = 0.81$, $p \ll 0.0001$), and so too for the Mandarin voices ($r_{188} = 0.69$, $p \ll 0.0001$).

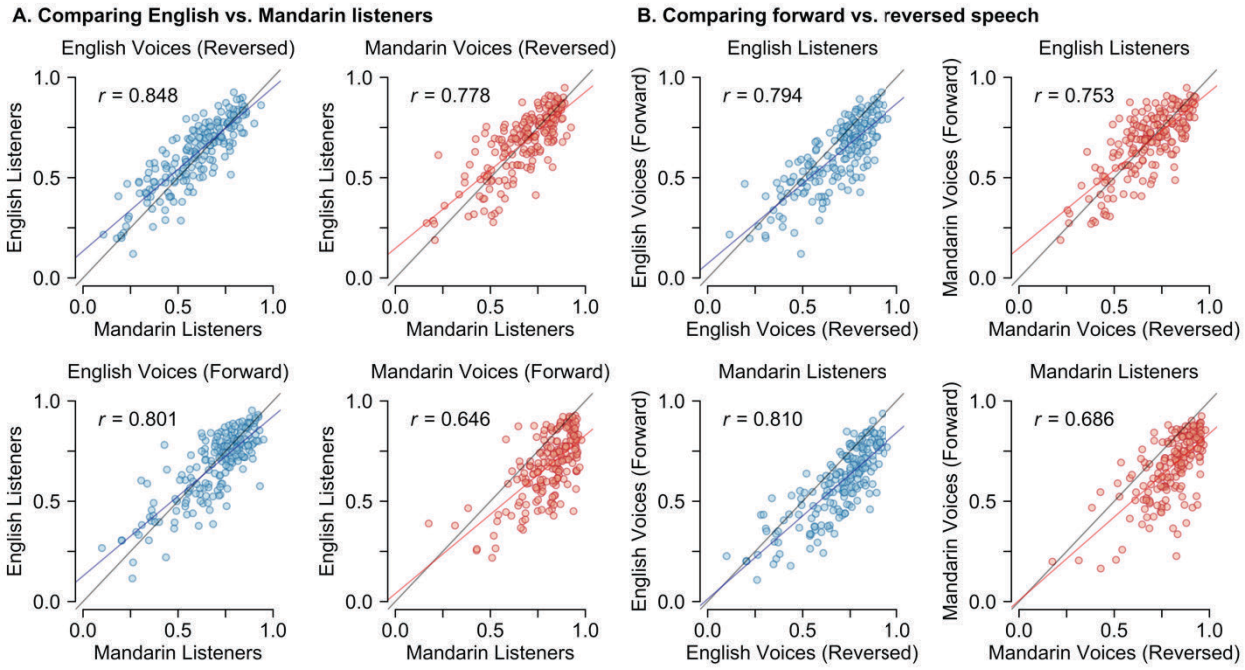

Figure 5: Correlation between dissimilarity judgments of voice pairs across listener groups and forward/time-reversed speech conditions. (A) Perceptual dissimilarity of voices was highly consistent across English and Mandarin listeners. Points indicate the mean perceived dissimilarity for each pair of voices for English (ordinate) and Mandarin (abscissa) listener groups. Both English and Mandarin listener groups tended to find the same voices more similar/dissimilar, as indicated by the high degree of correlation in these points. **(B)** Perceptual dissimilarity of voices was also highly consistent across forward and time-reversed speech, regardless of talker or listener language. Points indicate the mean perceived dissimilarity for each pair of voices for listeners who heard time-reversed (ordinate) or forward (abscissa) speech. The high degree of correlation for these points indicate that listeners found the same voices more similar/dissimilar regardless of whether the signal had been time-reversed or not.

D. Representational similarity analyses of speech acoustics and perceptual dissimilarity judgments

We next sought to identify which acoustic factors, if any, were related to listeners' perceptual dissimilarity judgments as a function of talker language, listener language, and time-reversal.

1. Acoustics of our Mandarin- and English-speaking talkers

Our two talker groups differed in a number of measures (**Table 4, Figure 6**). For instance, the English-speaking talkers on average had lower vocal pitch and smaller formant dispersion, perhaps suggesting smaller overall body size in our Mandarin-speaking sample (cf. Pisanski et al., 2014). Interestingly, the English-speaking talkers also tended to have more nonmodal voice quality, indicated by higher jitter and lower HNR values compared to the Mandarin talkers. Finally, English talkers tended to speak faster

than Mandarin talkers, perhaps owing to the presence of unstressed syllables and function words in the English recordings (IEEE, 1969) compared to the content-word heavy Mandarin sentences (Fu et al., 2011).

Table 4. Mean ± standard deviation and two-sample difference of acoustic measures from English- and Mandarin-speaking talkers.

	Group average* (mean ± s.d.)		**Group difference**		
Acoustic Feature	**English talkers**	**Mandarin talkers**	***t*(38) =**	***p* <**	**Cohen's *d* =**
f_0 mean (Hz)	212.35 ± 19.25	233.53 ± 29.02	-2.73	0.01	0.88
f_0 variation ($s/\bar{x}$)	0.152 ± 0.042	0.160 ± 0.025	-0.75	0.46	0.24
Jitter (%)	1.034 ± 0.152	0.760 ± 0.132	6.06	0.0001	1.97
HNR (dB)	13.20 ± 1.362	15.02 ± 1.708	-3.72	0.001	1.21
Speech rate (syl. / s)	5.305 ± 0.366	4.824 ± 0.449	3.71	0.001	1.20
Formant dispersion (Hz)	1101.4 ± 53.4	1198.2 ± 85.0	-7.99	0.0001	2.59

*n = 20 in each talker group

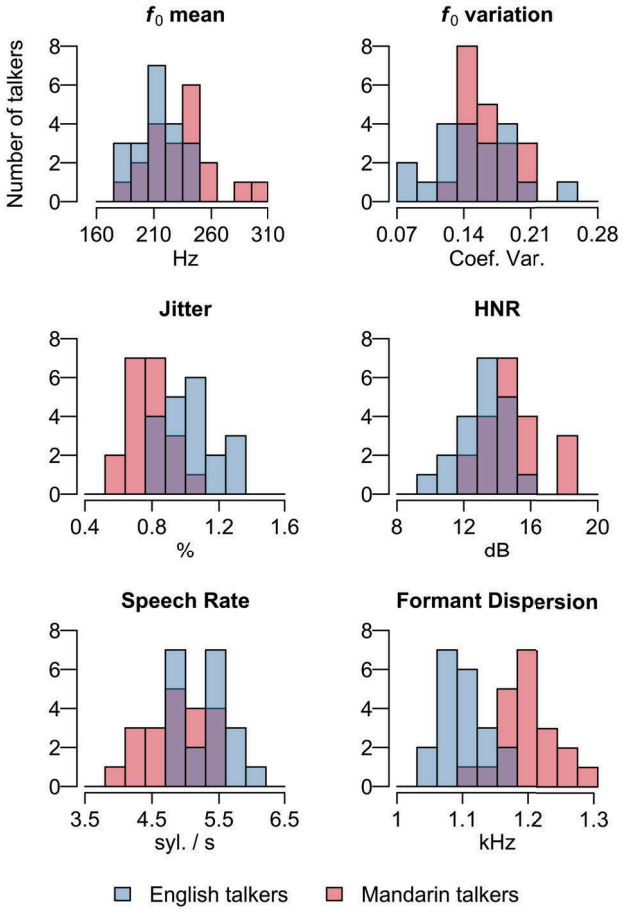

Figure 6: Acoustic features of English and Mandarin talkers. Histograms display the number of talkers falling within a particular range for each of the acoustic features measured. Means, standard deviations, and differences between groups are reported in **Table 1**.

2. Data analysis

Many of the acoustic features we measured are subject to nonlinear mapping between physical and perceptual space. Consequently, when applicable we scaled these values into the corresponding perceptual space for comparison to listeners' behavior: Talkers' f_0 and formant frequencies were converted to mels

(Stevens, Volkman, & Newman, 1937). Variation in f_0 was scaled with respect to mean f_0 using the coefficient of variation ($s/\bar{x}$). Perception of jitter and HNR are roughly linear over the range of values measured in our speakers (Hillenbrand, 1988), and so these values were not transformed. Dissimilarity judgment data were arcsin transformed prior to inclusion in the representational similarity analysis models (Studebaker, 1985).

Using a representational similarity analysis technique (Kriegeskorte, Mur, & Bandettini, 2008), we performed a backward stepwise regression on a linear mixed effects model that estimated the effect of the six acoustic measures (f_0 *mean,* f_0 *variation, jitter, HNR, speech rate, and formant dispersion*) on each listener's perceptual dissimilarity judgment on each within-language different-talker trial (i.e., all the native and foreign talker pairs, but not cross-language or same-talker pairs). The model additionally contained all two-, three-, and four-way interaction terms between these continuous measures and the categorical factors of *listener native language* (English, Mandarin), *talker language* (English, Mandarin), and *recording direction* (time-reversed, forward). The random effects structure included by-participant intercepts and slopes for the within-subject categorical fixed effect *talker language*, as well as random by-item intercepts for each talker pair. A deviation contrast coding scheme was applied to all categorical factors. Backward stepwise regression was performed using the function *step* in the package *lmerTest* for both fixed and random effects, with the criterion for factor inclusion of $\alpha = 0.05$.

3. Acoustic factors affecting perceptual dissimilarity judgments of voices

Following the backward stepwise regression analysis, a number of acoustic factors and their interactions with talker- and listener-specific linguistic factors had significant effects on listeners' perceptual dissimilarity judgments (**Table 5**). With respect to the continuous acoustic factors, there were significant overall effects of the difference in *mean* f_0, *HNR*, and *formant dispersion* between talkers in a pair on listeners' dissimilarity ratings, such that larger differences in these features tended to lead to higher dissimilarity ratings. Speech rate also played a role in the model, but only in the context of its interactions with categorical factors (see below). With respect to the categorical factors there were, like before, significant effects of *talker language* (with higher dissimilarity ratings overall for Mandarin than English talker pairs) and of *recording direction* (with higher dissimilarity ratings overall for forward than time-reversed recordings), but again no overall effect of *listener native language*.

The final representational similarity analysis model included no significant two-way interactions between acoustic differences and talker or listener language. There was a significant two-way *HNR* × *recording direction* interaction, indicating that differences between talkers' HNR had a larger influence on dissimilarity ratings of time-reversed speech than forward speech. The significant two-way interaction between the categorical variables of *talker language* and *listener language* recapitulates the effect seen in **Table 2** and **Figure 4A**. No other two-way interactions were significant.

The reduced model included only one significant three-way interaction involving differences in speech acoustics: *speech rate* × *talker language* × *listener language*, such that both Mandarin and English listeners tended to rate talker pairs in their native language as more dissimilar when the difference in speech rate was smaller, but English listeners more strongly exhibited the opposite pattern in their foreign language, rating Mandarin talker pairs as more dissimilar with larger differences in speech rate. The three-way *talker language* × *listener language* × *recording direction* interaction parallels that in **Table 2**.

Table 5. Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices.

Model term	*β*	*s.e.*	*t*	*df*	*p*-value
Δ f_0 mean (mel)	0.0024	0.0002	9.62	25370	≪ 0.0001
Δ HNR (dB)	0.0087	0.0029	3.01	30206	0.0027
Δ Speech rate (syl. / s)	-0.0111	0.0113	-0.99	29896	0.32
Δ Formant dispersion (mel)	0.0007	0.0003	2.64	30103	0.0083
*Talker language**	0.0654	0.0244	2.68	514	0.0075
Listener native language	-0.0609	0.0482	-1.26	79	0.21
Recording direction	-0.1422	0.0482	-2.95	79	0.0042
Δ Speech rate × *Talker language*	0.0075	0.0113	0.67	29879	0.51
Δ Speech rate × *Listener native language*	0.0029	0.0096	0.31	29909	0.76
Talker language × *Listener native language*	-0.0703	0.0125	-5.62	154	≪ 0.0001
Δ HNR × *Recording direction*	0.0061	0.0025	2.48	29880	0.013
Talker language × *Recording direction*	0.0117	0.0105	1.11	76	0.27
Listener native language × *Recording direction*	0.0565	0.0477	1.18	76	0.24
Δ Speech rate × *Talker language* × *Listener native language*	0.0210	0.0096	2.18	29904	0.029
Talker language × *Listener native language* × *Recording direction*	0.0359	0.0105	3.43	76	0.0010

*Categorical factors are shown in italics.

IV. DISCUSSION

A. Perceptual dissimilarity ratings

Listeners' perceptual dissimilarity judgments evinced a number of consistent patterns that transverse both listeners' native language background and manipulation via time-reversal. On average, listeners judged between-language pairs to be most dissimilar, suggesting that listeners may generally expect speech in different languages to come from different talkers, and that, even when time-reversed, the acoustic qualities of speech in English and Mandarin are sufficiently different to give the consistent impression that these recordings come from different talkers. This is further supported by the numerous significant differences in low-level acoustic features between the two talker groups – including mean vocal pitch, voice quality measures, formant dispersion, and speech rate (**Table 4**). Differences in the low-level acoustic measurements between the two talker groups appear to be reflected in listeners' heightened perceptual dissimilarity judgments across languages (**Fig. 3D**).

It is also important to note that, while differing significantly in mean dissimilarity, the distributions of dissimilarity ratings for cross-language voice pairs and same-language voice pairs are nonetheless highly overlapping (**Fig. 3C**). This indicates that mere differences in the language being spoken do not uniquely determine the extent to which pairs of voices will be heard as sounding similar – even for natural speech stimuli – and that the features giving rise to perceptual dissimilarity of voices must transcend the linguistic typology or content of speech. While these observations expand on those reported by Fleming and colleagues (2014), who also used English and Mandarin voices, future work remains needed to explore how these observations generalize to pairings of other languages, as well as how listeners might judge dissimilarity of same-talker / cross-language pairs, which were not included in the design of Fleming and colleagues or this replication (cf. Winters, Levi, & Pisoni, 2008).

Relatedly, listeners gave the lowest perceptual dissimilarity ratings to pairs of recordings coming from the same talker, regardless of talker language, listener language, or whether the recordings had been

time-reversed. This strongly suggests that, for any given speaker, the individuating acoustic features of their voice are largely robust to disruption by time-reversal and speech content, and are largely preserved across linguistic differences in listeners' experiences with voices. Moreover, the distributions of mean dissimilarity ratings for same-voice pairs and cross-voice pairs are essentially non-overlapping across listeners and languages (**Fig. 3C**), further suggesting that listeners' perception of the distinct acoustic features associated with a particular talker are, in fact, highly discriminative, even when asked to judge subjective similarity. That is, across recordings and even distortion via time-reversal, any given talker is, on average, much more likely to sound like herself than like any other talker, even talkers of the same language, and moreover even when that language is completely unfamiliar to listeners. This reveals that listeners are highly perceptually sensitive to the individuating acoustic features of voices, even while they are simultaneously challenged in their ability to learn to associate those features with a particular talker's identity (Perrachione, 2018; Perrachione et al., 2015). This observation adds compelling further evidence that the discrimination and identification of voices are, by and large, two fundamentally separate abilities (Van Lancker & Kreiman, 1987; Perrachione et al., 2014; Fecher & Johnson, 2018), which should raise caution in ascribing causal mechanisms to phenomena in talker identification from data derived from different tasks. For instance, the language-familiarity effect is widely attested in talker identification (e.g., Goggin et al., 1991; Perrachione & Wong, 2007; Bregman & Creel, 2014, Xie & Myers, 2015; *inter alia*), but it is not clear that we can use results from other tasks such as *talker discrimination* (Johnson et al., 2011; Wester, 2012) or *perceptual dissimilarity judgments* (Fleming et al., 2014) to identify the causal mechanisms behind superior native-language talker identification abilities. Effects from these other tasks are unlikely to yield dispositive evidence about the underlying cognitive or perceptual mechanisms at play in talker identification (Levi, 2018). This view is further endorsed by a recent study of the development of the language-familiarity effect, showing that even when listeners exhibit a robust native-language bias in talker identification, they may show no effect of language on talker discrimination (Fecher & Johnson, 2018).

Finally, unlike the prior report of Fleming and colleagues (2014), we did not observe a consistent difference in listeners' perceptual dissimilarity judgments for native- vs. foreign-language talker pairs. While this overall effect was newly found in natural speech recordings (albeit inconsistently between the two groups, see below), we failed to replicate the prior observation of higher dissimilarity ratings for recordings of time-reversed native-language talkers compared to time-reversed foreign-language ones. This result is inconsistent with the view that the phonological features giving rise to the language-familiarity effect are reliably present even in time-reversed speech. Instead, this result parallels the observation that talker identification from time-reversed voices is also less susceptible to the language-familiarity effect (Perrachione et al., 2015), suggesting instead that listeners' perceptual and mnemonic processing of time-reversed voices may be largely independent of any linguistic (i.e., phonological or lexical) features in the speech of talkers or language-specific representations in the minds of listeners. Because we failed to replicate the core finding of Fleming and colleagues (2014), in the following section, we explore the patterns of listener- and talker-language effects in perceptual dissimilarity judgments of voices in greater detail.

B. A language-familiarity effect in perceptual judgments of voice dissimilarity?

Although prior reports have suggested that listeners' perceptual dissimilarity judgments of voices reveal a phonological basis for the language-familiarity effect, there is no direct evidence for preservation

of language-specific phonological features in time-reversed speech. Our observation that time-reversed native-language voices are not consistently judged to be more dissimilar than foreign-language ones also calls this interpretation into question, particularly since the language-familiarity effect has otherwise been so widely replicated in talker identification tasks (as reviewed in Perrachione, 2018; Levi, 2018). How, then, might linguistic factors affect perceptual dissimilarity judgments of voices?

Native speakers of both Mandarin and English judged time-reversed Mandarin voices to be more dissimilar than time-reversed English voices (**Fig. 4A**). If language familiarity affects perceptual dissimilarity judgments and talker identification abilities in the same way, then this pattern of perceptual dissimilarity judgments by English-speaking listeners is unexpected and inconsistent with that hypothesis. For the time-forward voices, too, English-speaking listeners found English-speaking voices no more dissimilar than Mandarin-speaking ones, notwithstanding their widely-reported difficulty learning to identify Mandarin talkers (Perrachione & Wong, 2007; Perrachione et al., 2009; 2011; 2015; McLaughlin et al., 2015; Xie & Myers, 2015; Zarate et al., 2015; McLaughlin et al., 2019). Instead, this result suggests that perceptual dissimilarity judgments of voices may ultimately have more to do with acoustic differences in the voices of talkers than with linguistic differences in the minds of listeners.

Turning to our novel analysis of perceptual dissimilarity judgments of *same-talker* voice pairs, our results again do not suggest a language-familiarity effect in perceptual dissimilarity judgments. Mandarin- and English-speaking listeners do not reliably judge foreign-language, same-talker voice pairs as sounding more dissimilar than native-language, same-talker voice pairs – the pattern we would expect if listeners were more sensitive to the distinguishing features of voices in their native language. This pattern of results is observed in both time-reversed and natural recordings, where talker identity should have been more easily ascertained (Sheffert et al., 2002; Remez et al., 2007; Perrachione et al., 2014; Perrachione et al., 2015): neither English- nor Mandarin-speaking listeners appear to favor their native language, contrary to the predictions of a language-familiarity effect for voice dissimilarity judgments. This suggests that listeners are actually quite good at "telling together" foreign-language voices (e.g., Lavan et al., 2019a; 2019b), even when they otherwise struggle to learn to identify those same voices (e.g., McLaughlin et al., 2019).

Taken together, these results provide little evidence for a language-familiarity effect on perceptual dissimilarity judgments of voices. The expected pattern of results – that different-talker voice pairs will be more dissimilar, and that same-talker voice pairs will be more similar, in a listener's native language – is infrequently attested and the exact opposite pattern of results is sometimes found instead. Even where the pattern seems consistent with the hypothesis – for instance in the Mandarin listeners' judgments of greater perceptual dissimilarity for time-reversed Mandarin voices – the hypothesis is rejected by other data, namely the identical-but-unexpected pattern of judgment by English listeners. Instead, an alternative hypothesis – that perceptual dissimilarity judgments of voices depend more on the acoustic features of the voices themselves than on perceptual biases arising from listeners' long-term experiences – appears more tenable.

C. Consistency of perceived dissimilarity across listeners and comprehensibility

In the language-familiarity effect, listeners are not only less accurate learning to identify voices speaking a foreign language, they also exhibit different patterns of talker identification errors than native-language listeners (Perrachione & Wong, 2007). If listeners *identify* voices based on the same features they use when judging their perceptual dissimilarity, we would expect to find divergence between the

dissimilarity judgments of listeners of different backgrounds. Instead, we found these judgments to be remarkably consistent across listeners of different language backgrounds (**Fig. 5**). Pairs of talkers thought to sound more similar by Mandarin-speaking listeners were also thought to sound more similar by English-speaking listeners, regardless of whether those voices were speaking English or Mandarin, and regardless of whether those voices were forward or time-reversed. Indeed, perception of talker dissimilarity was also robust to time-reversal, with pairs of voices judged to sound more similar in time-forward speech also judged to sound more similar in time-reversed speech. Together, these results suggest that perceptual dissimilarity judgments of voices are likely to be made on the basis of acoustic features of speech and voice that are independent of speech comprehensibility (or even naturalness), and which are largely unaffected by the linguistic structure of speech. Correspondingly, we next explored the possibility that language-independent acoustic features form the basis for perceptual dissimilarity judgments of voices.

D. Acoustic features in perceptual dissimilarity judgments of voices

We measured the relationship between trial-by-trial differences in various acoustic features of voices and listeners' ratings of perceived dissimilarity of those voices. Across both listener groups, both talker languages, and both forward and time-reversed speech, differences in talkers' mean fundamental frequency were most strongly related to listeners' dissimilarity judgments. For a pair of recordings in which talkers exhibited greater differences in mean f_0, listeners were more likely to rate that pair of voices as sounding dissimilar, all other factors notwithstanding. Other acoustic factors were also related to listeners' judgments of talker dissimilarity, including HNR and formant dispersion. As acoustic measures, HNR may capture perceptual correlates of voice quality related to periodicity in the glottal cycle, while formant dispersion indexes vocal tract length – both individuating acoustic features of the voice that should be preserved across time reversal and readily salient in both languages.

If language familiarity affects listeners' dissimilarity judgments of voices through differential familiarity with the pattern of low-level acoustic features in their native language vs. a foreign language, we would expect to see significant interactions between listeners' native language and the various acoustic measures affecting perceptual dissimilarity judgments. In fact, we observed only one such interaction, relating not to a low-level acoustic property of talkers' voices, but rather to speech rate: both Mandarin and English listeners tended to actively discount differences in speech rate as an index of dissimilarity in their native languages; however, English listeners appeared to change their strategy and rely on this as a measure of dissimilarity when listening to Mandarin. Interestingly, this difference did not appear to be affected by time-reversal suggesting it is not related to comprehensibility – that is, that Mandarin listeners could understand the natural speech in both languages, whereas English listeners would find only natural English speech comprehensible.

The only other acoustic factor that interacted with our linguistic manipulations was HNR, which played a larger role in dissimilarity judgments for time-reversed voices than time-forward ones. The HNR measurement should capture perceptual qualities related to the periodicity of the vocal source, which should be preserved under time-reversal, even while other phonetic and phonological relationships are obfuscated by that manipulation. That HNR played a greater role in perceptual dissimilarity judgments under time-reversal in both languages may suggest that listeners increased their reliance on this cue in the absence of other phonetic features on which they would usually base their judgments. This may be related to the observation that judgments of dissimilarity were more extreme for natural recordings compared to

time-reversed ones, where listeners would have found fewer familiar phonetic features from which to make their judgment.

E. What can we learn about perception of voices from subjective judgments?

If perceptual dissimilarity judgments of voices are unlikely to reveal the cognitive bases of the language-familiarity effect in talker identification, then what else can we learn about human voice processing from this method? Our experience conducting this study suggests that this kind of paradigm has several fundamental limitations that may constrain its utility in answering theoretical questions about voice processing.

First, listeners appear predisposed to make judgments at the extreme ends of the response range (**Fig. 3B**). Listeners are, overall, very good at discriminating whether two voices are the same or different (e.g., Wester, 2012), and, consequently, their judgments of dissimilarity often take a binary form: listeners' responses primarily encode whether they believe two stimuli are the same voice or not, even when asked to judge voice dissimilarity on a continuous scale. Thus, this paradigm has limited sensitivity for identifying either cognitive or acoustic factors involved in the perception of voices.

Second, listeners' dissimilarity judgments may not be based on the same features for every trial. For example, listeners may largely rely on mean f_0 in dissimilarity judgments, but then lean on differences in other features like speech rate or formant dispersion only when differences in mean f_0 are particularly small. Listeners may also exhibit nonlinear correspondences between acoustic features and their dissimilarity judgments, such that a linear model – even accounting for perceptual warping of acoustic space as we have done – will fail to adequately model how listeners map perceptual space to subjective dissimilarity for a complex acoustic stimulus like a voice (e.g., Krieman & Sidtis, 2011).

Third, the interpretability of experimental paradigms involving time reversal of speech pose serious problems because of the extent to which this manipulation is unnatural, both in terms of listeners' experience with such stimuli and the ecological validity of the resulting acoustic features themselves. For example, ecologically invalid acoustic features of time-reversed speech include the physiologically-impossible reversed shape of the glottal waveform, the unnatural trajectory of formant transitions and their rates of change, the unnatural trajectory of the amplitude contour, the presence of uncommon or phonotactically impermissible phonetic transition probabilities, and so on. In the present work, for instance, we found that listeners appeared to rely primarily on salient, non-articulatory acoustic cues (e.g., mean pitch) when judging perceived dissimilarity for speech. Reliance on such cues, which convey only a fraction of talker identity (e.g., Remez, Fellowes, & Rubin, 1997; Perrachione et al., 2014), may suggest why listeners tend to perform so poorly when learning to identify talkers from time-reversed speech (Sheffert et al., 2002) and why they may fail to exhibit a language-familiarity effect from such stimuli (Perrachione et al., 2015).

Fourth, just because listeners believe two voices sound similar – or dissimilar – does not necessarily mean they will be more or less likely to confuse those voices during discrimination, recognition, or identification tasks. It is possible that the acoustic cues that listeners prioritize when rating voice dissimilarity differ from those that underlie the holistic auditory gestalt that gives rise to their perception of voice identity (e.g., Kreiman & Sidtis, 2011). It remains to be seen whether subjective judgments of voice dissimilarity correspond to objective measurements of listeners' talker discrimination or talker identification skills.

Finally, it is worth noting that the relatively large number of cross-language pairs may have introduced bias in listeners' dissimilarity judgments. Because cross-language pairs were consistently more different than within-language pairs – both in terms of their acoustics and listeners' judgments of them – their relative over-representation in this design may have "anchored" listeners' expectations about the magnitude of differences they could expect, leading them to report greater within-language similarity than they otherwise might have. Likewise, following Fleming and colleagues (2014), we did not include cross-language / same-talker pairs in our design. How listeners would behave when rating the subjective dissimilarity of recordings of the same talker speaking different languages remains to be seen. (A change in language between talker familiarization and later recognition test does appear to introduce a reduction in recognition accuracy, revealing that some cues to talker identity are likely to be language-specific; Winters, Levi, & Pisoni, 2008.)

F. Potential alternatives for more effective measurements of talker dissimilarity judgments

If subjective perceptual dissimilarity ratings are encumbered by limitations on their generalizability and ecological validity, and if other laboratory tasks, such as talker identification or discrimination, are also limited in their ability to tell us about ecological voice *perception* due to their disproportionate demands on long-term memory or low-level acoustic analysis, respectively, then what other options may be available to develop a more ecological understanding of voice dissimilarity judgments in future studies? A key approach will be to treat voices in the laboratory the same way we do psychologically: holistically. Rather than force listeners into paradigms where they seem to make decisions based on salient, low-level acoustic features, future paradigms can require listeners to make judgments based on the vocal gestalt. For instance, a recently developed tasks asks listeners to indicate *when* a talker changes (not just *whether*), allowing the sensitive measure of response time to reveal subtle differences in how listeners detect differences between voices (Sharma et al., 2019). Combining ratings of perceived voice dissimilarity with a change detection task can reveal whether and how dissimilarity measures are related to ecological voice perception behaviors, such as switching attention to a new talker in a mixed-talker setting.

In applying a more holistic approach to studying perceived dissimilarity, listeners could be presented with two pairs of voices and be required to indicate which pair is more dissimilar – thus reducing their tendency to provide an extreme dissimilarity rating for any pair of voices separately. These patterns of responses can be analyzed using approaches adopted from psychological models of *comparative judgment* (Thurstone, 1927) and then related to stimulus acoustics. Beyond a winner-takes-all approach, the relative dissimilarity of two pairs of voices can be analyzed using *magnitude estimation* methods borrowed from psychophysics (e.g., Poulton, 1968), which have seen similar utility in other fields of linguistics (e.g., Bard, Robertson, & Sorace, 1996) where listeners are also otherwise biased towards the extreme end of a bounded rating scale. Finally, researchers can undertake targeted examinations of acoustic features, in isolation or combination, that they hypothesize underlie dissimilarity judgments of voices through sophisticated acoustic resynthesis as available in software packages like STRAIGHT (Kawahara et al., 2008). Hypothesis-driven approaches to acoustic manipulation have already done much to inform the psychological foundations of voice perception (e.g., Latinus & Belin, 2011a; Latinus et al., 2013); future work will be able to use similar approaches to examine how acoustic and linguistic factors – including differences between the languages spoken by talkers and listeners – interact in listeners' perception of voices.

V. CONCLUSIONS

(i) Listeners' perceptual dissimilarity judgments of voices provide weak and inconsistent evidence of a language-familiarity effect in voice processing, especially compared to the effect sizes reported in prior literature using talker identification or talker discrimination tasks. (ii) Overall, listeners of different language backgrounds tend to make perceptual judgments of voice similarity that are more similar than different, regardless of whether they are listening to voices in their native language and regardless of whether the voices have been rendered incomprehensible by time-reversal. (iii) Of the acoustic features analyzed here, mean f_0 tended to have the greatest effect on listeners' judgments of voice dissimilarity regardless of the language spoken by talkers or listeners; however, in general, listeners' judgments of voice dissimilarity do not appear to map neatly onto isolated acoustic features. (iv) These results ultimately suggest that the language-familiarity effect in voice processing is more likely to be due to linguistic or mnemonic bases than perceptual ones.

VI. OPEN-SOURCE DATASET

All audio recordings used as stimuli, experimental paradigms, behavioral data, and analysis scripts are available via this project's online archive: https://open.bu.edu/handle/2144/16460

VII. ACKNOWLEDGMENTS

We thank Deirdre McLaughlin, Gabriel Cler, Yaminah Carter, Sara Dougherty, Jennifer Golditch, Andrea Chang, Cecilia Cheng, and Sung-Joo Lim for their assistance collecting the data and discussing the analyses. Research reported in this article was supported by the NIDCD of the National Institutes of Health under award number R03DC014045 and by a NARSAD Young Investigator Award from the Brain and Behavior Research Foundation to T.P.

NOTES

[1]As noted during peer review, a counterbalancing error affected the number of trials per condition for one participant, with three too many same-identity trials, one too few cross-language trials, and two too few native-language trials. Unique items were nonetheless heard on all trials, and correct total number of trials were heard. This error affected 0.006% of the data collected.

VIII. REFERENCES

Bard, E.G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, 72, 32-68.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, 74, 110-120.

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

Davis, S.B. (1981). Acoustical characteristics of normal and pathological voices. *ASHA Rep.*, 11, 97-115.

Davison, D. S. (1991). An acoustic study of so-called creaky voice in Tianjin Mandarin. UCLA Working Papers in Phonetics, 78:50-57.

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

Fitch, W.T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *Journal of the Acoustical Society of America*, 102, 1213-1222.

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, 13795-13798. doi: 10.1073/pnas.1401383111

Fu, Q.-J., Zhu, M., & Wang, X. (2011). Development and validation of the Mandarin speech perception test.

Journal of the Acoustical Society of America, 129(6), EL267–EL273.

Ganugapati, D., & Theodore, R. M. (2019). Structured phonetic variation facilitates talker identification. *Journal of the Acoustical Society of America*, EL469.

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

Hillenbrand, J., Getty, L.A., Clark, M.J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099-3111.

IEEE. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*, 17(3), 225–246.

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

Kang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *Journal of the Acoustical Society of America*, 142, 1693-1706.

Karnell, M.P., Melton, S.D., Childes, J.M., Coleman, T.C., Dailey, S.A., & Hoffman, H.T. (2007). Reliability of clinician-based (GRBAS and CAPE-V) and patient-based (V-RQOL and IPVI) documentation of voice disorders. *Journal of Voice,* 21, 576-590.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H.Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation, *Proc. ICASSP 2008, Las Vegas,* 3933-3936.

Keating, P. & Esposito, C. (2007). Linguistic voice quality. UCLA Working Papers in Phonetics #105, pp. 85-91

Kempster, G.B., Gerratt, B.R., Verdolini Abbott, K., Barkmeier-Kramer, J., & Hillman, R.E. (2009). Consensus Auditory-Perceptual Evaluation of Voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18, 124–132.

Kreiman, J. (1982). Perception of sentence and paragraph boundaries in natural conversation. *Journal of Phonetics*, 10, 163-175.

Kreiman, J., Gerratt, B.R., & ud Dowla Khan, S. (2010). Effects of native language on perception of voice quality. *Journal of Phonetics*, 38, 588-593.

Kreiman, J. & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception.* Wiley-Blackwell.

Kreiman, J., Vanlancker-Sidtis, D., & Gerratt, B. (2005). Perception of voice quality, in D.B. Pisoni & R.E. Remez (Eds.), *The Handbook of Speech Perception*, Blackwell Publishing: Malden, MA.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. doi: 10.3389/neuro.06.004.2008

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

Latinus, M., & Belin, P. (2011b). Human voice perception. *Current Biology*, 21(4), R143–R145. https://doi.org/10.1016/j.cub.2010.12.033

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, 23, 1075-1080. https://doi.org/10.1016/j.cub.2013.04.055

Lavan, N., Burston, L.F.K., & Garrido, L. (2019a). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, 110, 576-593.

Lavan, N., Burton, A.M., Scott, S.K., & McGettigan, C. (2019b). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26, 90-102.

Lavan, N., Merriman, S.E., Ladwa, P., Burston, L.F.K., Knight, S., & McGettigan, C. (2018). Please sort these sounds into 2 identities: Effects of task instructions on performance invoice sorting studies. https://psyarxiv.com/yu34f/

Levi, S.V. (2018). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science*. https://doi.org/10.1002/wcs.1483

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

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, and Psychophysics*, 81, 1088-1107.

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

Mok, P. (2008) On the syllable-timing of Cantonese and Beijing Mandarin. In Proceedings of the 8th Phonetics Conference of China (PCC 2008) and the International Symposium on Phonetic Frontiers (ISPF 2008). Beijing.

Peirce, J.W. (2007). PsychoPy - Psychophysics software in python. *Journal of Neuroscience Methods*, 162(1), 8–13.

Perrachione, T.K. (2018). Recognizing speakers across languages, in S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception*, Oxford: Oxford University Press.

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

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

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

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

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

Pisanski, K., Fraccaro, P.J., Tigue, C.C., et al., (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, 95, 89-99.

Poulton, E.C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, 69, 1-19.

Remez, R.E., Fellowes, J.M., & Nagel, D.S. (2007). On the perception of similarity among talkers. *Journal of the Acoustical Society of America*, 122, 3688–3696.

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

Schweinberger, S.R. & Zaske, R. (2018). Perceiving speaker identity from the voice. In Fruhholz & Belin, Eds., *The Oxford Handbook of Voice Perception*, Oxford University Press.

Sharma, N.K., Ganesh, S., Ganapathy, S., & Holt, L.L. (2019). Talker change detection: A comparison of human and machine performance. *Journal of the Acoustical Society of America*, 145, 131-142.

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

Shih, C. L. (1988). Tone and intonation in Mandarin. Work Papers of the Cornell Phonetic Laboratory, 3, 83-109.

Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, 20, 171-186.

Slifka, J. (2007). Irregular phonation and its preferred role as a cue to silence in phonological systems. 16th International Congress of Phonetic Sciences, Saarbrucken, August 2007.

Stevens, S.S., Volkmann, J., & Newman, E.B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8, 185–190.

Studebaker, G.A. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research*, 28, 455-462.

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

Van Lancker, D. & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, 25, 829-834.

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

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

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

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, 419-432.

Xue, S.A., Hao, G.J.P., & Mayo, R. (2006). Volumetric measurements of vocal tracts for male speakers from different races. *Clinical Linguistics and Phonetics,* 20, 691-702.

Yang, B.G. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics*, 24, 245-261.

Yang, R.-K. (2011). The phonation factor in the categorical perception of Mandarin tones. *17th International Congress of Phonetic Sciences* (Hong Kong, August 2011).

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, 11475.

Zraick, R.I., Kempster, G.B., Connor, N.P., Thibeault, S., Klaben, B.K., Bursac, Z., Thrush, C.R., & Glaze, L.E. (2011). Establishing validity of the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V). *American Journal of Speech-Language Pathology*, 20, 14-22.