n

Fleming, D., Giordano, B. L., Caldara, R., and Belin, P. (2014) *A language-familiarity effect for speaker discrimination without comprehension.* Proceedings of the National Academy of Sciences of the United States of America, 111 (38). pp. 13795-13798. ISSN 0027-8424.

Copyright © 2014 The Authors

http://eprints.gla.ac.uk/97494/

Deposited on: 26 September 2014

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

A language-familiarity effect for speaker discrimination without comprehension

David Fleming[a,1], Bruno L. Giordano[a], Roberto Caldara[b], and Pascal Belin[a,c,d]

[a]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QB, United Kingdom; [b]Department of Psychology, University of Fribourg, 1700 Fribourg, Switzerland; [c]Département de Psychologie, Université de Montréal, Montréal, QC, Canada H2V 2S9; and [d]Institut des Neurosciences de La Timone, Unite Mixte de Recherche 7289, Centre National de la Recherche Scientifique, Université Aix-Marseille, 13005 Marseille, France

Edited by E. Anne Cutler, University of Western Sydney, Penrith South, New South Wales, Australia, and approved August 8, 2014 (received for review January 23, 2014)

The influence of language familiarity upon speaker identification is well established, to such an extent that it has been argued that "Human voice recognition depends on language ability" [Perrachione TK, Del Tufo SN, Gabrieli JDE (2011) *Science* 333(6042):595]. However, 7-mo-old infants discriminate speakers of their mother tongue better than they do foreign speakers [Johnson EK, Westrek E, Nazzi T, Cutler A (2011) *Dev Sci* 14(5):1002–1011] despite their limited speech comprehension abilities, suggesting that speaker discrimination may rely on familiarity with the sound structure of one's native language rather than the ability to comprehend speech. To test this hypothesis, we asked Chinese and English adult participants to rate speaker dissimilarity in pairs of sentences in English or Mandarin that were first time-reversed to render them unintelligible. Even in these conditions a language-familiarity effect was observed: Both Chinese and English listeners rated pairs of native-language speakers as more dissimilar than foreign-language speakers, despite their inability to understand the material. Our data indicate that the language familiarity effect is not based on comprehension but rather on familiarity with the phonology of one's native language. This effect may stem from a mechanism analogous to the "other-race" effect in face recognition.

voice perception | unintelligible speech | dissimilarity ratings

The human voice carries linguistic information as well as paralinguistic information about a speaker's identity, and normal listeners possess abilities to extract both types of information. The neuro-cognitive mechanisms underlying speech comprehension and speaker recognition are dissociable, as evidenced by cases of both patients with receptive aphasia (impaired speech comprehension but preserved speaker recognition) and patients with phonagnosia (impaired speaker recognition but preserved speech comprehension) (1–5), as well as by differences in the cortical networks engaged by the two abilities (6–13). However, speech and voice identity processing also interact to a considerable degree. Speech recognition is influenced by speaker variability and familiarity: listeners better understand and remember speech spoken by familiar speakers (14–17). Conversely, speaker identification is influenced by language familiarity: listeners are typically poorer at identifying speakers of a foreign language. This so-called "Language-Familiarity Effect" (LFE) has been demonstrated across a diverse range of languages (18–22) and is behaviorally robust, persisting even after several days of training (23).

A crucial, unresolved point of debate is whether the LFE depends upon linguistic mechanisms involved in speech comprehension, or rather reflects the greater familiarity with the phonological structure of one's own language without necessarily requiring an understanding of the linguistic message. On the one hand, evidence from dyslexic participants, whose phonological processing abilities are impaired (24), supports the importance of linguistic processing for general speaker identification abilities: English-speaking dyslexic participants do not show the LFE, (i.e., better memory for English-speaking than Chinese-speaking voices) shown by normal participants (25). On the other hand, a LFE is already apparent in infants before they can fully comprehend speech: 7-mo-olds notice a speaker change in their native language but not in an unfamiliar language (26). Although results from dyslexic participants suggest a specific link between the LFE and "language ability" (25), results from infants (26) suggest that experience with the phonology of the maternal language, rather than comprehension, may underpin the LFE. If this is the case, then the enhanced individuation of own-language speakers observed in 7-mo-olds should be observed in normal adult participants, even for unintelligible speech.

Here we tested this hypothesis by comparing dissimilarity ratings of own- and different-language speakers with time-reversed speech stimuli. Note that reversing speech disrupts intelligibility, but preserves "considerable phonetic information" (27) as well as sufficient indexical information to enable listeners to recognize voices (28–30). We collected speaker dissimilarity ratings from Chinese and English listeners for all pairwise combinations of a set of Mandarin ($n = 20$) and English ($n = 20$) time-reversed speech clips, and compared these dissimilarity ratings between groups of speakers and listeners. If the LFE is based primarily on language comprehension, then we should observe no interlanguage difference in discrimination performance, as time-reversal rendered all stimuli unintelligible. Conversely, familiarity with a language's characteristic phonological structure may suffice to engender a LFE for speaker discrimination, even without comprehension. Mandarin Chinese is a tonal language, whereas English is stress-based; as such, a Mandarin speaker and an English speaker may differ in terms of the language structure elements that they use to differentiate speaking voices. For example, Mandarin and English

Significance

A recent report [Perrachione TK, Del Tufo SN, Gabrieli JDE (2011) *Science* 333(6042):595] shows that dyslexic individuals do not show an advantage in the recognition of speakers of their own language compared with speakers of a foreign language (the well established "language-familiarity" effect) as typical subjects do, a finding that Perrachione et al. interpreted as evidence that "Human voice recognition depends on language ability." Here, we refine this notion by providing, to our knowledge, the first evidence of a language-familiarity effect in adult cross-cultural listeners with speech rendered unintelligible by time reversal. This result shows that the phonological aspects of language ability only, even without comprehension, can influence speaker discrimination.

Author contributions: D.F., B.L.G., R.C., and P.B. designed research; D.F. performed research; D.F. and B.L.G. analyzed data; and D.F., B.L.G., R.C., and P.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. Email: d.fleming.1@research.gla.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1401383111/-/DCSupplemental.

differ in speaking fundamental frequency (31–33) and phonemic inventories: Mandarin features around 1,300 syllables, whereas English uses around 15,000 (34); the languages have very little consonant overlap; and English features a high frequency and variety of consonant clusters, whereas Mandarin has no consonant clusters (35, 36). Time-reversal preserves the formant structure (in a "mirrored" form; ref. 27) of many phonemes and their mean fundamental frequency, and, given that these features may differ across the two languages in natural speech, and that they are relatively well-preserved upon reversal, then native speakers of both languages may still be sensitive to these differences even where intelligibility is disrupted. If sensitivity to such differences drives a LFE then each group should show higher dissimilarity ratings for pairs of voices speaking their native language than for pairs speaking the other language

Results

All possible paired combinations of voices were presented to listeners who recorded their dissimilarity ratings via a computerized visual analog scale, ranging from 0 to 1 (where a rating of 0 corresponded to maximum perceived similarity and 1 to maximum perceived dissimilarity). Fig. 1*A* shows the dissimilarity matrix averaged across English and Chinese participants, where rows/columns 1–20 correspond to native voices and rows/columns 21–40 to foreign voices. Participants rated four types of pair: same-identity trials (where the same speaker was heard twice within a pairing), foreign–foreign trials, native–native trials, and native–foreign trials. No sentence clip was uttered twice within a pair. As shown in Fig. 1*A*, interlanguage pairs (where presentations consisted of one native and one foreign voice) were rated as more dissimilar than all other pairs, as reflected by the overall red color (high dissimilarity) of the upper right and lower left submatrices. Fig. 1*B* illustrates the differences between each rating condition (same-identity mean = 0.16 ± 0.02 SE; foreign–foreign mean = 0.59 ± 0.02 SE; native–native mean = 0.62 ± 0.02 SE; native–foreign mean = 0.71 ± 0.02 SE). Each participant's mean ratings for each trial type were submitted to a repeated-measures ANOVA, which revealed a significant effect of pair type [$F(3, 39) = 314.2$, $P < 0.001$, $\eta^2_{partial} = 0.89$]. Post hoc tests also revealed significant differences for all pairwise comparisons of trial type (all P values < 0.02).

Crucially, when taking participant groupings into account, both Chinese- and English-speaking listeners produced higher average dissimilarity ratings for native-language voice pairs than for nonnative-language pairs (Chinese: native mean = 0.62 ± 0.03 SE, nonnative mean = 0.60 ± 0.03 SE; English: native mean = 0.61 ± 0.02 SE, nonnative mean = 0.57 ± 0.02 SE) (Fig. 1 *B* and *C*). We submitted these ratings to a 2 × 2 mixed-measures ANOVA, with listener language and speaker language as the between- and within-group (repeated) measures, respectively. A significant interaction between speaker and listener's language was observed, indicating that native-language dissimilarity ratings were higher, regardless of the language group of the listener [$F(1, 38) = 11.13$, $P = 0.002$, $\eta^2_{partial} = 0.23$]. The main effects of both listener and speaker language were not significant (P values > 0.2), suggesting that there were no statistical differences in rating behavior between groups and that both sets of voices elicited similar rating behavior. Paired t tests confirmed our prediction that both listener groups rated own-language pairs as more dissimilar than different-language pairs (Chinese-speaking participants: Chinese > English [$t(19) = 2.57$, $P = 0.02$, Cohen's $d = 0.17$]; English-speaking participants: English > Chinese [$t(19) = 2.36$, $P = 0.03$, Cohen's $d = 0.41$]). To investigate the robustness of these results, we computed bootstrapped 95% confidence intervals of the native > foreign

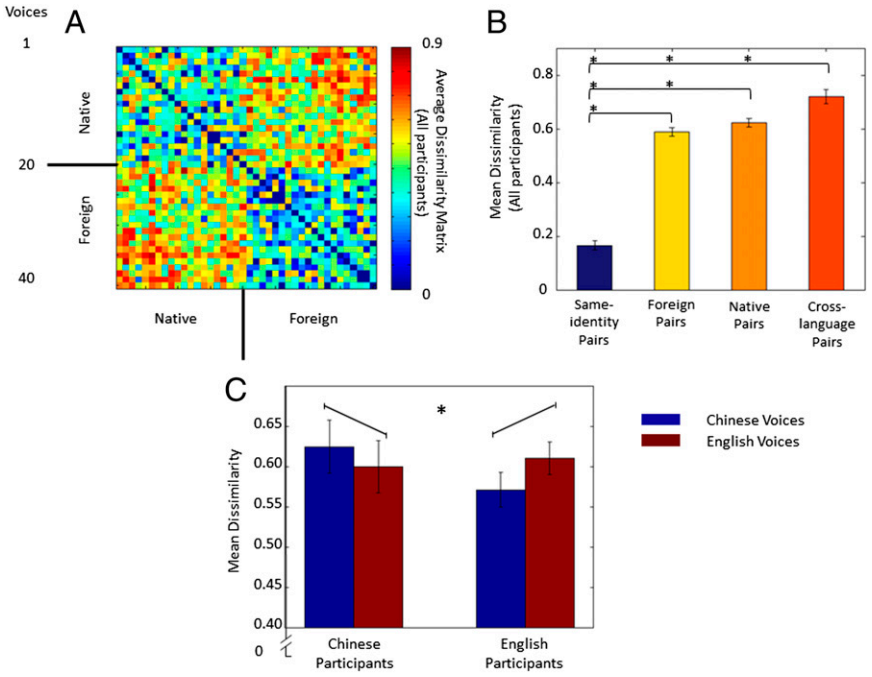

Fig. 1. Speaker dissimilarity ratings for pairs of Mandarin and English time-reversed sentences. (*A*) Matrix of dissimilarity ratings averaged across all participants in both listener groups ($N_{Chinese\ Listeners} = 20$; $N_{English\ Listeners} = 20$): individual participants' dissimilarity matrices are in a standardized arrangement, so that rows and columns 1–20 (top and left) represent native voices (Mandarin for Chinese listeners, English for English listeners), whereas rows and columns 21–40 represent foreign language voices, regardless of listener group. The color scale indicates group-average dissimilarity ratings. (*B*) Average dissimilarity ratings for the four different types of pairs. Cross-language pairs were rated as most dissimilar. Within the same-language pairs, crucially, native-language pairs were rated as more dissimilar than foreign-language pairs, even though all sentences were unintelligible. (*C*) Listener × speaker interaction: both participant groups record higher average dissimilarity ratings for native-language vs. foreign-language speaker pairs. Error bars indicate the SEM. All asterisks denote $P < 0.05$.

difference for each group, and for all participants taken together. We sampled participants' difference scores with replacement (10,000 iterations) and derived separate confidence intervals for each group (CI for Chinese participants: [0.007–0.04]; English: [0.005–0.07]; combined groups: [0.01–0.05]). As none of these confidence intervals contained zero, the observed effects may be considered reliable.

Discussion

We investigated whether the LFE in adults requires comprehension of the linguistic message. We found that listeners rated pairs of speakers of their own language as more dissimilar on average than pairs of speakers of a different language, even though all stimuli were rendered unintelligible by time-reversal. This result implies that the LFE is not rooted in language comprehension per se, but rather is based on familiarity with the acoustical fingerprint of one's language, in a manner analogous to the "Other-Race Effect" (ORE) in face recognition.

Participants were presented with pairs of time-reversed sentences spoken by different speakers and were asked to judge how dissimilar the speakers were. Time-reversal was chosen because it is a simple procedure that compromises intelligibility while preserving some of the information present in the natural speech signal. For example, time-reversal disrupts the temporal attributes of speech segments, such as onsets and decays, and reverses pitch curves. Conversely, reversed speech is identical to natural speech in amplitude, duration, and mean fundamental frequency. Furthermore, the formant transition structure of many phonemes (e.g., fricatives and long vowels) is approximately mirrored in the reversed signal, and important indexical cues to speaker identity are also retained. In sum, this remaining information can enable high intersubject agreement in phoneme transcription tasks (27), and can be used by the listener to aid speaker recognition (28–30). Our participants were unable to extract any meaning from the stimuli, yet they showed reliable differences in their identity dissimilarity ratings. The most salient difference was between the different-language pairs (i.e., consisting of one sentence in English and one sentence in Mandarin) and the same-language pairs: The listeners reliably rated pairs of different-language speakers as more dissimilar than pairs where the language was consistent across identities (either both speakers in English or both in Mandarin), clearly visible in the dissimilarity matrix in Fig. 1*A* as red and green submatrices. This result confirms that subjects were able to use acoustical information in the time-reversed sentences and were sensitive to overall acoustical differences between the two languages (Table S1 and *SI Methods*).

Crucially for our hypothesis, listeners also rated pairs of speakers of their own language as more dissimilar than pairs of speakers of the other language. The effect is highly significant and apparent as an interaction when ratings are split by speaker and listener group in Fig. 1*C*. This effect is not driven by one subject group, as there is no main effect of subject group on overall ratings and the own-language effect is significant for each subject group individually. Nor is it explained by one of the sets of stimuli as the effect of speaker language on the ratings was not significant either. However, despite the absence of a main effect of listener group, the native-language bias appears to be stronger in the English listener group compared with the Chinese, reflected in the differences in effect sizes. This result may be explained by the fact that our Chinese participants had been resident in the United Kingdom for 9 mo on average at the time of testing, and had considerable functional experience with the English language. It has been demonstrated, for example, that nonnative speaker identification performance improves over several days' worth of training (23).

Our results provide the first evidence, to our knowledge, of a LFE in adult participants in the absence of speech comprehension. These findings extend the results of Johnson et al. (26), who observed a similar effect in 7-mo-old infants: In both cases, subjects had a limited understanding of the stimuli, yet they were more sensitive to identity differences in native-language pairs compared with nonnative pairs. Interestingly, however, the infants in Johnson et al.'s (26) study did not show a discrimination bias for reversed native speech compared with reversed foreign speech, as our adult listeners did. The infants' comparatively lower experience with the phonology of their native language may account for this; specifically, whereas 7 mo of exposure may be sufficient to enable differentiation of native speakers uttering normal speech, it may be insufficient for the kind of fine-grained differentiation required under alien processing conditions, as in the case of reversed speech. Indeed, even school-aged children may not display adult-like performance in speaker recognition tasks (37), suggesting that they cannot use the information available in an utterance as effectively as an adult listener despite their substantial experience of their native phonology and their greater exposure to different voices, compared with infants. Therefore, it may be that infants do not yet possess the ability to extract information from an unintelligible speech signal to aid speaker discrimination and recognition, in ways that adults can, as shown in our discrimination results and previous recognition results (28–30).

Thus, our findings refine Perrachione et al.'s (25) view that "human voice recognition depends on language ability" by supporting the notion that phonological processing is the key aspect of language ability which facilitates speaker individuation, but adding that comprehension of the spoken message is not necessary for such individuation. Their findings suggest that impaired voice recognition in dyslexics may be driven by their known deficits in phonological processing (24), whereas our results show that the limited phonological information and indexical cues preserved after time-reversal are sufficient to allow listeners to differentiate speakers. Moreover, extended exposure with the particular distribution of acoustical features characteristic of their own language allowed our participants to perceptually "zoom-in" on speakers of that language, resulting in higher native-language dissimilarity ratings, even when both native and nonnative speech content was unintelligible.

These findings draw an interesting parallel with an analogous effect in another sensory modality: the ORE in face recognition. The ORE is the well known phenomenon where observers are typically poorer at discriminating and recognizing faces from a different racial group compared with their own (for a review, see ref. 38). One influential account of the ORE suggests that individual faces are represented as points in a multidimensional space whose dimensions are shaped by perceptual experience and code for diagnostic features (39, 40). Own-race faces, with which an observer has more experience, become distributed more diffusely about the origin of the space (i.e., the average or prototypical face). Other-race faces, as a result of a different statistical distribution of features, are encoded in a less efficient manner due to a reliance on diagnostic dimensions for individuation optimized for own-race faces. Other-race faces therefore mistakenly appear more similar to one-another to the observer and this confusability between faces underpins the impairment in other-race recognition performance. Indeed, this model has found support at the behavioral, computational (41) and neurophysiological levels (42–44). An analogous model could be invoked to account for our results and those of Johnson et al. (26). One could conceive of a similar "voice space" where voices are encoded as points based on experience with indexical and linguistic attributes. Indeed, the behavioral and physiological relevance of such a voice space model has already been demonstrated (45). Speakers of one's native language would, in such a framework, be represented in a more distributed manner, resulting in higher interspeaker discriminability than for speakers of other languages to which the subject has had less exposure and are therefore represented in

a less differentiated, more compact manner. Such a model, while acknowledging that comprehension can modulate speaker identification, would be consistent with the many noted similarities between cerebral face and voice processing (46).

Methods

Participants. Twenty Mandarin-speaking Chinese (8 female, mean age = 23.7, SD = 2.58) and 20 native English-speaking UK participants (10 female; mean age = 24.25, SD = 3.01) were recruited. Chinese participants' average duration of UK residency was 9.35 mo (SD = 7.34) and all had attained a minimum score of 6.5 on the IELTS test of English as a foreign language, or a comparable score on an equivalent test. English-speaking participants reported no experience with Mandarin Chinese. All participants were right-handed and reported no history of hearing difficulties or pathology. Participants gave written, informed consent for their involvement and received a monetary reward. The experiment was approved by the Ethical Committee of the University of Glasgow's College of Science and Engineering.

Stimuli. Testing stimuli were drawn from a pool of 400 clips of 40 female speakers (20 native English-speaking and 20 native Mandarin-speaking) reading 10 sentences (Open Speech Repository, 2005). Recordings were digitized at a 16-bit/44.1-kHz sampling rate and cut into individual sentences. Full, sentence-length stimuli were subsequently time-reversed, standardized to duration of 1,250 ms (from original onset), and normalized for RMS amplitude. The use of time-reversed clips of English and Mandarin speech ensured that stimuli are of equal intelligibility to both participant groups and therefore largely eliminated the influence of spoken language comprehension upon participants' behavior. Stimuli were edited using Adobe Audition 2 (Adobe Systems) and MATLAB 7.10 (R2010a).

Procedure. Testing took place within an anechoic cabin, where participants were seated at a desktop PC. The experiment was programmed in MATLAB 7.5 (R2007a). On every trial, participants heard a pair of voices and were instructed to rate the likelihood that both voice clips had been produced by the same speaker, using a visual analog scale where 0/far-left corresponded to "Same" and 1/far-right corresponded to "Different." Participants were advised to use the full extent of the scale to record their responses and were permitted to replay a trial as many times as they felt necessary before responding. This procedure was repeated for all possible paired combinations of voice identity, yielding a total of 820 pairs ($40 \times 39/2 + 40$ same-identity pairs). The assignment of sentence to speaker was randomized across identities, ensuring that no two voices in a pair ever produced the same sentence clip and that each participant received a unique series of sentence-to-speaker pairings, in addition to a unique identity pairing order. The self-paced experiment took participants ~2 h to complete, including an optional break when they had reached trial 411 (i.e., halfway through the experiment). Participants had received previous exposure to the voice stimuli in this experiment through their participation in an earlier functional Magnetic Resonance Imaging (fMRI) experiment, the results of which are not discussed here.

ACKNOWLEDGMENTS. P.B. was supported by the Biotechnology and Biological Sciences Research Council (BB/J003654/1).

1. Assal G, Aubert C, Buttet J (1981) Asymétrie cérébrale et reconnaissance de la voix. *Rev Neurol (Paris)* 137(4):255–268.
2. Van Lancker D, Kreiman J (1987) Voice discrimination and recognition are separate abilities. *Neuropsychologia* 25(5):829–834.
3. Van Lancker DR, Cummings JL, Kreiman J, Dobkin BH (1988) Phonagnosia: A dissociation between familiar and unfamiliar voices. *Cortex* 24(2):195–209.
4. Van Lancker DR, Kreiman J, Cummings J (1989) Voice perception deficits: Neuroanatomical correlates of phonagnosia. *J Clin Exp Neuropsychol* 11(5):665–674.
5. Garrido L, et al. (2009) Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia* 47(1):123–131.
6. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. *Nature* 403(6767):309–312.
7. Belin P, Zatorre RJ, Ahad P (2002) Human temporal-lobe response to vocal sounds. *Brain Res Cogn Brain Res* 13(1):17–26.
8. Belin P, Zatorre RJ (2003) Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport* 14(16):2105–2109.
9. Poeppel D (2003) The analysis of speech in different temporal integration windows: Cerebral lateralization as "asymmetric sampling in time". *Speech Commun* 41:245–255.
10. von Kriegstein K, Eger E, Kleinschmidt A, Giraud AL (2003) Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res* 17(1):48–55.
11. Kriegstein KV, Giraud A-L (2004) Distinct functional substrates along the right superior temporal sulcus for the processing of voices. *Neuroimage* 22(2):948–955.
12. Poeppel D, Idsardi WJ, van Wassenhove V (2008) Speech perception at the interface of neurobiology and linguistics. *Philos Trans R Soc Lond B Biol Sci* 363(1493):1071–1086.
13. McGettigan C, Scott SK (2012) Cortical asymmetries in speech perception: What's wrong, what's right and what's left? *Trends Cogn Sci* 16(5):269–276.
14. Martin CS, Mullennix JW, Pisoni DB, Summers WV (1989) Effects of talker variability on recall of spoken word lists. *J Exp Psychol Learn Mem Cogn* 15(4):676–684.
15. Mullennix JW, Pisoni DB, Martin CS (1989) Some effects of talker variability on spoken word recognition. *J Acoust Soc Am* 85(1):365–378.
16. Pisoni DB (1993) Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Commun* 13(1-2): 109–125.
17. Nygaard LC, Pisoni DB (1998) Talker-specific learning in speech perception. *Percept Psychophys* 60(3):355–376.
18. Thompson CP (1987) A language effect in voice identification. *Appl Cogn Psychol* 1: 121–131.
19. Goggin JP, Thompson CP, Strube G, Simental LR (1991) The role of language familiarity in voice identification. *Mem Cognit* 19(5):448–458.
20. Köster O, Schillert NO (1997) Different influences of the native language of a listener on speaker recognition. *Forensic Linguist.* 4:18–28.
21. Winters SJ, Levi SV, Pisoni DB (2008) Identification and discrimination of bilingual talkers across languages. *J Acoust Soc Am* 123(6):4524–4538.
22. Perrachione TK, Pierrehumbert JB, Wong PCM (2009) Differential neural contributions to native- and foreign-language talker identification. *J Exp Psychol Hum Percept Perform* 35(6):1950–1960.
23. Perrachione TK, Wong PCM (2007) Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia* 45(8):1899–1910.
24. Gabrieli JDE (2009) Dyslexia: A new synergy between education and cognitive neuroscience. *Science* 325(5938):280–283.
25. Perrachione TK, Del Tufo SN, Gabrieli JDE (2011) Human voice recognition depends on language ability. *Science* 333(6042):595.
26. Johnson EK, Westrek E, Nazzi T, Cutler A (2011) Infant ability to tell voices apart rests on language experience. *Dev Sci* 14(5):1002–1011.
27. Binder JR, et al. (2000) Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex* 10(5):512–528.
28. Bricker PD, Pruzansky S (1966) Effects of stimulus content and duration on talker identification. *J Acoust Soc Am* 40(6):1441–1449.
29. Van Lancker D, Kreiman J, Emmorey K (1985) Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *J Phonetics* 13:19–38.
30. Sheffert SM, Pisoni DB, Fellowes JM, Remez RE (2002) Learning to recognize talkers from natural, sinewave, and reversed speech samples. *J Exp Psychol Hum Percept Perform* 28(6):1447–1469.
31. Eady SJ (1982) Differences in the F0 patterns of speech: Tone language versus stress language. *Lang Speech* 25:29–42.
32. Mang EA (2001) Cross-language Comparison of Preschool Children's Vocal Fundamental Frequency in Speech and Song Production. *Res Stud Music Educ* 16:4–14.
33. Keating P, Kuo G (2012) Comparison of speaking fundamental frequency in English and Mandarin. *J Acoust Soc Am* 132(2):1050–1060.
34. Shu H, Anderson RC (1999) Learning to read Chinese: The development of metalinguistic awareness. *Reading Chinese Script: A Cognitive Analysis*, eds Wang J, Inhoff A, Chen H (Lawrence Erlbaum Associates, Publishers, New Jersey).
35. Yeong SHM, Rickard Liow SJ (2012) Development of phonological awareness in English-Mandarin bilinguals: A comparison of English-L1 and Mandarin-L1 kindergarten children. *J Exp Child Psychol* 112(2):111–126.
36. Duanmu S (2000) *The Phonology of Standard Chinese* (Oxford Press, New York).
37. Mann VA, Diamond R, Carey S (1979) Development of voice recognition: Parallels with face recognition. *J Exp Child Psychol* 27(1):153–165.
38. Meissner CA, Brigham JC (2001) Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychol Public Policy Law* 7:3–35.
39. Valentine T (1991) A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A* 43(2):161–204.
40. Valentine T, Endo M (1992) Towards an exemplar model of face processing: The effects of race and distinctiveness. *Q J Exp Psychol A* 44(4):671–703.
41. Caldara R, Abdi H (2006) Simulating the 'other-race' effect with autoassociative neural networks: Further evidence in favor of the face-space model. *Perception* 35(5): 659–670.
42. Vizioli L, Rousselet GA, Caldara R (2010) Neural repetition suppression to identity is abolished by other-race faces. *Proc Natl Acad Sci USA* 107(46):20081–20086.
43. Vizioli L (2012) *Clarifying the neurophysiological basis of the other-race effect.* Ph. D. Thesis (University of Glasgow, Glasgow, United Kingdom).
44. Brosch T, Bar-David E, Phelps EA (2013) Implicit race bias decreases the similarity of neural representations of black and white faces. *Psychol Sci* 24(2):160–166.
45. Latinus M, McAleer P, Bestelmeyer PEG, Belin P (2013) Norm-based coding of voice identity in human auditory cortex. *Curr Biol* 23(12):1075–1080.
46. Yovel G, Belin P (2013) A unified coding strategy for processing faces and voices. *Trends Cogn Sci* 17(6):263–271.