Learning to recognize unfamiliar voices: the role of language familiarity and music experience

Micah R. Bregman (mbregman@cogsci.ucsd.edu)
Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515
La Jolla, CA 92093 USA

Sarah C. Creel (creel@cogsci.ucsd.edu)
Department of Cognitive Science, UC San Diego, 9500 Gilman Dr. M/S 0515
La Jolla, CA 92093 USA

Abstract

Speech not only transmits semantic information through words and syntax, but also provides cues to a talker's identity. Differences in a listener's ability to recognize voices can be attributed to their language background, and in rare cases voice recognition can be selectively damaged in neurological patients. In this study we investigated a group of Korean-English bilinguals and non-Korean speakers' ability to learn to recognize unfamiliar Korean and English talkers by voice, and to generalize to utterances not heard during training. We observed an interaction between language background and stimulus language for speed of learning, however generalization performance indicated no such interaction when compared to baseline performance. Bilinguals' performance recognizing English (but not Korean) voices, was predicted by the age they learned English. We also observed that individuals who actively participated in music production exhibited significantly faster task learning than those who did not produce music. This study indicates that language background has a gradient effect on voice learning among bilinguals, and that non-linguistic auditory processing differences, such as music perception, impact voice identification.

Keywords: speech perception; music perception; voice; voice identification; individual differences; bilingualism

Introduction

Speech is generally studied primarily for its ability to communicate semantic meaning from one individual to another. Many complex animal communication systems such as birdsong, however, evolved primarily to communicate more basic information, providing cues that other conspecific listeners use to evaluate fitness and individual identity. Any comprehensive understanding of the evolutionary origins of speech and language will draw both upon the role communication signals play in transmitting semantic meaning, as well as their role in providing cues to identity.

Human speech contains many acoustic cues that listeners use to recognize, for example, a talker's age, gender, emotional state, or even their identity. Collectively, these elements of the speech signal are known as "indexical cues". Voice recognition, or talker identification, is an important aspect of speech perception, and one that has been relatively little studied. Although often considered separate from the core speech perception system (some neuroimaging results support this perspective, e.g. Belin, Fecteau, & Bédard, 2004), several studies suggest that talker-specific acoustic cues are intertwined with speech recognition. For example, listeners are better able to understand speech from familiar talkers than unfamiliar ones (Nygaard & Pisoni, 1998).

While several studies have characterized severe disability in voice identification, few have attempted to investigate differences among individuals' abilities to recognize voices, although the existence of dramatic individual differences has been noted for many years (Pollack, Pickett, & Sumby, 1954). In clinical cases, voice recognition can be lost completely in individuals with a neuropsychological disorder known as phonagnosia (Van Lancker, Kreiman, & Cummings, 1989). In a pioneering study, Goggin, Thompson, Strube, & Simental (1991) demonstrated that monolingual English speakers were better able to identify the voices of English-German bilinguals when listening to those individuals speak English than when they spoke German. This suggested that, despite many shared acoustic features (both English and German stimuli shared the acoustic features imparted by a particular talker's vocal tract), the listener's language background had a strong impact on their ability to recognize the voices. This study suggested that differences in phonological processing that arise from linguistic knowledge are important in voice recognition.

Goggin et al. (1991) observed no difference in performance on a voice recognition task for English-Spanish bilinguals when tested on English vs. Spanish speaking voices. They suggested that bilinguals might have equal ability recognizing voices from either language since they have extensive phonological knowledge of both. Bilinguals, however, are heterogeneous in their language background, and it may be the case that late learners, or those dominant in one of their languages do exhibit the voice identification deficits identified in monolinguals.

A recent study demonstrated that differences in phonological processing *within* a language can also affect voice identification. Individuals with dyslexia are significantly impaired in their ability to recognize voices relative to controls, but only in their native language (Perrachione, Del Tufo, & Gabrieli, 2011). This result implies that individual differences in phonological

processing, even among those who share a language background, can dramatically impact listeners' abilities to recognize voices.

Outside clinical populations, what other differences might affect voice recognition accuracy? One possibility is music experience. Extensive musical training may benefit the neural encoding of speech by driving brain networks involved in both speech and music perception to function with higher precision than normally necessary for speech perception alone (Patel, 2011). In fact, musicians have been demonstrated to outperform non-musicians on speech perception tasks, including enhanced perception of speech in noise (Parbery-Clark, Skoe, Lam, & Kraus, 2009) as well as enhanced second language phonological ability in bilinguals (Slevc & Miyake, 2006). Do differences in music background or music perception affect voice recognition ability?

In this study, we investigated these questions in a group of Korean-English bilinguals and a second group of non-Korean speakers. We examined whether differences in language and music background, as well as individual differences in music perception and phonological working memory, affected participants' abilities to learn to recognize a set of unfamiliar voices. We also tested recognition of novel sentences spoken by these voices.

Methods

Participants

We tested 48 participants, 22 of whom were bilingual, and spoke Korean and English fluently. The remaining 26 participants had no background or experience with Korean. All Korean-English bilingual participants learned Korean as their first language or in parallel with English, and learned English between 1-17 years of age (mean=7.1 years). All subjects studied at UC San Diego and received course credit for participation. All procedures were part of a protocol approved by the UC San Diego Human Research Protections Program.

Stimuli

We recorded 15 Korean sentences spoken by each of four female native Korean speakers and 15 English sentences spoken by four female native American English speakers. English sentences were selected from the SPIN sentence set. All chosen sentences were high predictability, e.g. "He caught the fish in his net" (Kalikow, Stevens, & Elliott, 1977). Korean sentences were simple, high predictability, and of similar syllabic length to the English sentences, written by a native Korean speaker, e.g. "공책을 집에 놓고 왔다" ("Gongchek eul jibeh nohgo watda," "I left the notebook at home"). Recordings were made in a sound isolated recording booth, and each monaural recording was trimmed to begin at sentence onset and normalized to a mean of 70dB.

Procedure

Voice Learning Task Participants learned to associate 20 training stimuli (5 sentences x 4 voices) with one of four cartoon objects, which differed in both shape and color. Each cartoon object represented a single talker. We chose cartoon objects rather than faces to control for differences in face discriminability across participants. To initiate a trial, participants clicked a cross in the center of the screen. On each trial, audio playback began simultaneously with the display of the two cartoon objects, one on the left and one on the right, equidistant from the center cross. During each training trial, participants clicked one of the two objects with the computer mouse and after clicking, the correct object remained on the screen to provide feedback until they made a second confirmation click.

Training blocks of 60 trials each were presented (with stimuli randomized within each block) until participants reached 85% correct—that is, they chose the target object on at least 51 of 60 trials in a single block (chance=50%). After reaching criterion, participants completed two test blocks, each with 120 trials. During test blocks, no feedback was provided and the screen was blank after making a response. Test blocks contained 60 trials encompassing the 20 training stimuli, as well as 60 trials containing 5 novel sentences produced by the 4 learned voices. The second test block contained 60 trials of the 20 stimuli learned during training and an additional 5 novel sentences. After completing the training and testing process for one language, participants completed the process in the other language (English or Korean). The language of the first block (Korean or English), the cartoon objects associated with each voice, and the positions of the two images on the screen on each trial were counterbalanced across subjects.

Behavioral assessments In addition to completing the voice learning task, participants completed assessments to identify individual differences in language and music background and perception. They completed a questionnaire describing their music training, including formal training and current performance activity. To assess their dominant language, bilingual subjects completed a bilingual dominance survey (BDS; Dunn & Fox Tree, 2009) and a picture naming task assessing lexical inventory in English and Korean (modified from Gollan, Weissberger, Runnqvist, Montoya, & Cera, 2011). All participants completed the pitch contour subtest from the Montreal Battery for the Evaluation of Amusia (MBEA) to measure differences in music perception ability (Peretz, Champod, & Hyde, 2003). During the MBEA test, participants heard 2 example melody pairs followed by 31 test melody pairs. For each pair, they provided a same/different judgment. All melody pairs had the same melodic contour and there were no out-of-key notes, making it a fairly subtle change. Each participant's score was recorded as the number of correct responses (observed range = 12-30, mean = 23.5).

For the Korean-English bilingual participants, language dominance measured using the BDS ranged from -15 (English dominant) to 20 (Korean dominant) and averaged -

0.22. Performance on the lexical naming task ranged from -27 to 18, with a mean of -9.48. These bilingual dominance measures were highly correlated (r=0.78), and both BDS and MiNT scores were highly correlated with the age English was learned (r=0.92 and r=0.75, respectively).

Phonological working memory was estimated by measuring each participant's digit span. Digit span has been used as an index of phonological working memory in many experiments (Baddeley & Hitch, 1977). Participants heard a series of 16 audio recordings with a female voice reading random sequences of English digits at a rate of 1 digit per second. Two sequences for each length were presented, in order, from 2-9 digits. After each recording, participants repeated the numbers they had heard. Scores were recorded as the number of sequences correctly repeated, with a maximum score of 16 (observed range = 7-15, mean = 10.7). Digit spans did not differ between language groups (Welch's $t(45.95)=0.83$, $p=0.41$).

Results

Language familiarity predicts learning speed

Previous research suggests that familiarity with a language is predictive of performance on voice identification tasks. However, its role predicting learning rate for unfamiliar voices has not been explicitly tested. We contrasted 22 Korean-English bilinguals with 26 listeners who did not speak Korean. We measured the number of blocks required to reach a criterion of 85% correct within a single block. A 3-way mixed model ANOVA (Figure 1) with Participant Language (English-only, Korean-English; between-participants), Talker Language (English, Korean; within-participants) and block order (English first vs. Korean first; between-participants) revealed no significant main effects of participant language background ($F(1, 44)=3.19$, $p=0.08$), stimulus language ($F(1, 44)=0.44$, $p=0.51$), or block order ($F(1, 44)=1.09$, $p=0.30$). However, there was a strong interaction between stimulus language and language background ($F(1, 44)=24.02$, $p<0.0001$).

Individually, Korean-English bilingual participants were faster to learn Korean talkers ($M=1.9$ training blocks) than English talkers ($M=3.5$ blocks; paired t-test $t(21)=-3.03$, $p=0.006$). Similarly, English-speaking participants learned English voices ($M=2.5$ blocks) faster than Korean voices ($M=4.5$ blocks; paired t-test $t(25)=4.14$, $p=0.0003$). No other interactions were statistically significant (all Fs<0.08, ps>0.78). Together, these data show that differences in learning rates are present as a function of language background.

We then looked at participants' maximum accuracy on training trials. Although trained to reach a criterion of 85% correct in a block, some participants achieved higher accuracy then others. Again we observed an interaction between language background and stimulus language (Figure 2a) in the maximum accuracy reached. A 2-way mixed ANOVA indicates no main effects of language background ($F(1, 46)=2.08$, $p=0.16$) or stimulus language ($F(1, 46)=1.11$, $p=0.30$), but a strong interaction ($F(1, 46)=15.51$, $p=0.0003$).

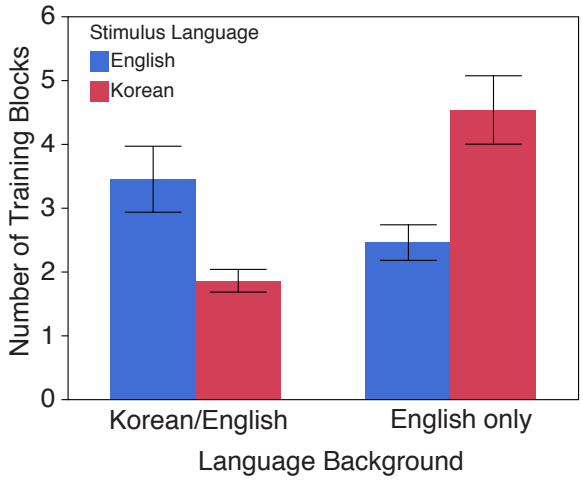

Figure 1: Korean-English bilinguals required fewer training blocks to reach 85% correct recognizing Korean speaking voices (red bars) than English speaking voices (blue bars). Non-Korean speakers show the opposite effect. Bars indicate mean number of training blocks ± s.e.

However, we observed no difference in performance between training and generalization test trials in the 40 participants who reached 85% correct after a maximum of 9 training blocks. For each of these participants, we calculated a "generalization penalty" by subtracting the proportion of correct responses to novel tokens of learned talkers with the proportion of correct responses to trained talkers. All stimuli were interleaved and collected in the same test block. We computed a 2-way mixed model ANOVA predicting participant's generalization penalty using language background (between participants) and stimulus language (within participants) as factors (Figure 2b). We observed no main effect of language background (Korean-English vs. English-only; between participants, $F(1, 39)=2.45$, $p=0.13$), no main effect of stimulus language (within participants, $F(1, 39)=1.72$, $p=0.20$) and no interaction between language background and stimulus language ($F(1, 39)=0.17$, $p=0.68$). While language background appears to be important for *learning* to distinguish unfamiliar voices, it does not appear to constrain generalizing to new utterances after the voices have been learned, at least within the short retention period required in this experiment.

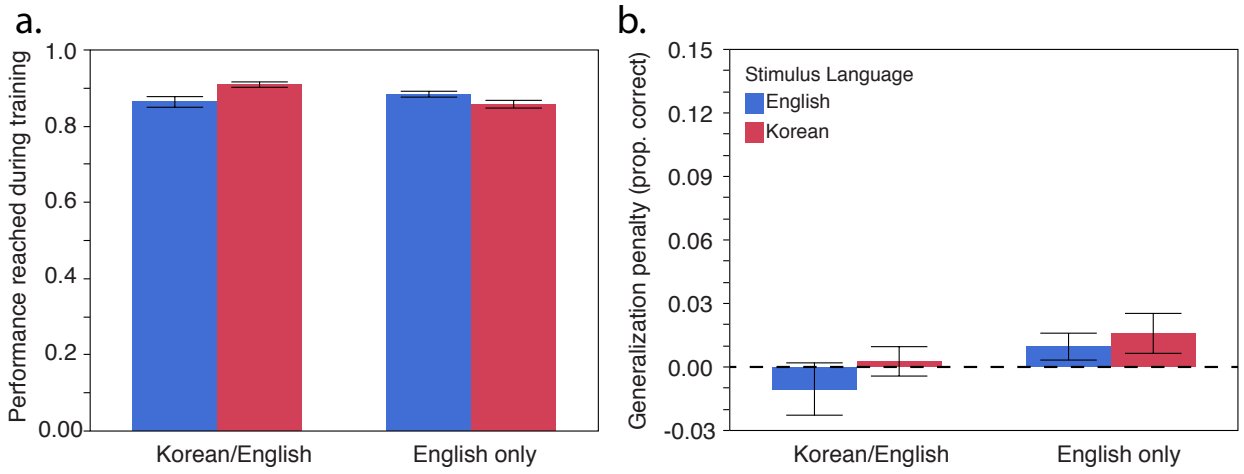

Figure 2: (a) Korean-English bilinguals were slightly more accurate at identifying the correct voice on novel sentences for Korean stimuli (red bars) than for English stimuli (blue bars). Non-Korean speakers show the opposite effect. (b) There were no generalization differences between groups

Bilinguals' age of L2 acquisition predicts learning speed in L2, but not L1.

We further explored whether individual differences in age of learning English or relative dominance of English or Korean were predictive of task performance among the bilingual subjects. To do so, we computed the correlation between age of English onset (which was the second language for all bilingual participants) with their voice learning rate. Among Korean-English bilinguals, blocks to criterion on English talkers was positively correlated with the age they began learning English (Figure 3a, $r(20)=0.62$, $p=0.002$), while it is uncorrelated for Korean-language stimuli ($r(20)=0.24$, $p=0.28$).

We then separated Korean-English bilingual participants into two groups based on a median split of acquisition age: those who learned English at or before 5 years old (early learners, $n=12$, mean age=3.3 years, mean BDS=-7.8, mean MiNT=-15.6) and those who learned after 5 years old (late learners, $n=10$, mean age=10.7 years, mean BDS=6.9, mean MiNT=-4.3). We then conducted a 2-way mixed model ANOVA with factors of Participant Language (between participants; English-only, early-English Bilingual, late-English Bilingual) and Talker Language (within participants). There was a main effect of language background ($F(2, 45)=4.73$, $p=0.014$), no main effect of stimulus language ($F(1, 45)=1.31$, $p=0.26$) and an interaction between language background and stimulus language ($F(2, 45)=15.91$, $p<0.0001$). This interaction resulted from three different patterns of talker learning. Early-learning bilinguals did not differ in their acquisition rate for Korean and English stimuli (paired $t(11)=-1.74$, $p=0.11$). However, late-English-learning bilinguals learned

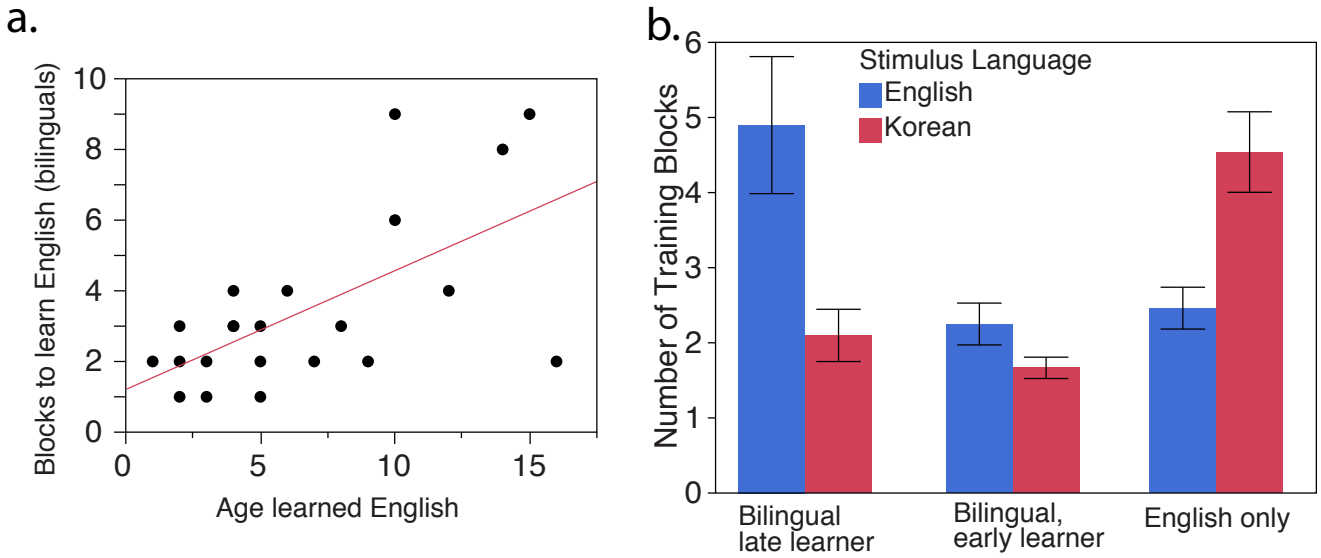

Figure 3: (a) The number of blocks to learn English voices was correlated ($r=0.62$, $p=0.002$) with the age Korean-English bilinguals learned English. (b) Number of training blocks to reach criterion of 85% on each stimulus language for Korean-English bilinguals who learned English late ($n=10$), early ($n=12$) and English-only speakers ($n=26$). Each bar represents the mean number of training blocks (60 trials/block) ± s.e.

Korean stimuli faster than English stimuli (paired $t(9)=-2.87$, $p=0.018$), and, as reported above, non-Korean speakers learned English stimuli faster than Korean stimuli.

Taken together, these results are consistent with prior work suggesting that phonological processing is an important element of voice recognition. Our result extends previous work by demonstrating a *gradient* effect of bilingualism. Rather than showing similar patterns of behavior in both languages, age of acquisition is an important predictor of performance recognizing voices in L2, but not L1.

Music experience predicts learning rate

We collected several behavioral measures of individual differences in auditory perception from our participants (see methods). Our hypothesis was that, since differences in individuals' language profiles (e.g. language familiarity, dyslexia) contribute to differences in voice learning, we might also observe differences among participants due to individual differences in auditory processing that are not strictly linguistic: pitch perception, music background, and music perception ability. We report the correlations between each of these measures and three performance measures: learning rate, generalization performance, and pitch shifted generalization performance (Table 1).

Several previous studies have identified perceptual advantages for individuals with extensive musical training. In particular, musicians have shown better brainstem encoding of pitch (Wong, Skoe, Russo, Dees, & Kraus, 2007), and high musical ability is associated with better second language phonology (Slevc & Miyake, 2006). Is musical experience important for learning to recognize voices?

Table 1. Correlations between music measures and voice recognition

					Generalization	
	Years Training	MBEA Score	Tone Thres.	*Learning rate (blocks)*	*Unshifted*	*Shifted*
Years Training	1.000	0.10	-0.02	**-0.42**	0.029	-0.095
MBEA Score		1.000	-0.26	-0.19	-0.048	-0.199
Tone Thres.			1.000	0.13	0.101	0.208
Learning rate				1.000	-0.202	0.246

We measured musical perceptual ability with the melody contour subtest of the MBEA (Peretz, Champod, & Hyde, 2003), and a pitch discrimination threshold task. Pitch difference threshold and MBEA did not correlate significantly with voice learning or generalization ability. However, measures of musical activity did show a relationship to voice learning rate. Participants who were currently active in producing music at least 1 hour per week when the experiment was conducted (n=11; musical training averaged 12.0 years, range 6-22 years) learned to recognize voices on average in fewer training blocks than those who were not active musicians (n=37; who had less musical training, averaging 5.2 years, range 0-27; Welch's $t(34.23)=-2.52$, $p=0.017$). This difference seems to have been driven by musicians' more rapid learning for voices speaking the subject's non-dominant language. When tested on the non-dominant language (Korean for non-Korean speakers, English for Korean-English bilinguals), musicians learned faster than non-musicians (mean=2.71 blocks vs. 4.62 blocks, Welch's $t(44.28)=-3.07$, $p=0.004$). However, when learning to recognize voices in their dominant language, we observed no effect of music background (mean=2.00 blocks for musicians vs. 2.40 blocks for non-musicians, Welch's $t(17.02)=-0.59$, $p=0.56$).

As there are multiple ways of assessing music experience, we also considered the effect of years of musical training (this did not overlap completely with current musical practice). Years of training correlated negatively with average number of training blocks to reach criterion ($r(46)=-0.42$, $p=0.0036$). Again, the relationship to music training is driven by the non-dominant language ($r(46)=-0.40$, $p=0.006$); musical training was not significantly correlated with learning rate for voices in the dominant language ($r(46)=-0.22$, $p=0.13$).

Discussion

Previous studies demonstrated that individual differences in phonological processing due to language background and dyslexia are important predictors of voice identification ability. The results of the current study extend these findings in a few important respects. In both adults and infants, knowledge of a language improves ability to recognize voices in that language (Goggin et al., 1991; Johnson, Westrek, Nazzi & Cutler, 2011; Perrachione et al., 2011). We extended this work by investigating both monolinguals and bilinguals, and looking at the bilingual participant's language dominance. Not only did we find a crossover interaction between listeners' native-language backgrounds and talkers' language, but we also found that early second-language acquisition facilitated talker learning without loss in performance on the first language. This acquisition effect—if viewed as such—is particularly interesting because it mimics acquisition of phonology: as age of acquisition increases, receptive and productive phonology are less native-like (Flege et al., 2006; Oh et al., 2011).

We also observed significantly faster voice learning for participants with more extensive musical training, particularly those actively involved in music production. This could be associated with changes in auditory encoding that have been observed among musicians that give rise to differences in pitch, music and speech perception. Our result extends this area of research, suggesting that not only is speech comprehension enhanced, but perception of indexical features in the speech signal may be enhanced as well. The effect of music experience appeared only to apply

to participants' learning to recognize voices in a less familiar language. We point out, however, that this study does not actually manipulate music training, so we cannot assert that it *causes* improvement in learning to recognize voices. Perhaps some third variable—inherent or learned individual differences in auditory perception—confers benefits to both voice recognition and music production.

Further work is also needed to identify whether the kinds of individual differences that give rise to enhanced voice recognition also extend to other indexical cues. Are individuals who performed better on individual recognition tasks also more sensitive to acoustic cues such as a talker's emotional state, age, or gender?

We explored how language experience and non-linguistic factors contributed to talker identification in two different languages. Native-language talkers were learned faster than second-language or unfamiliar-language talkers, and among bilinguals, earlier L2 acquisition predicted faster learning. Further, some measures of music experience predicted faster learning in the less-familiar language. Our work suggests a role for early language learning, or at least extent of exposure, in talker identification. This is consistent with a tight linkage between language processing and talker identification, which presents an interesting puzzle given the evidence of specialized neural mechanisms for speech recognition and talker identification.

Acknowledgements

MRB was supported by the Kavli Institute for Brain and Mind at UC San Diego, and SCC was supported by an NSF CAREER Award (BCS-1057080). We would like to acknowledge the help of undergraduate research assistants Shawn Cho and Hye Young Lee who were instrumental in developing Korean language stimuli and collecting data on participant's language background.

References

Baddeley, A. D., & Hitch, G. J. (1977). Working Memory. The psychology of learning and motivation: advances in research and theory.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-35.

Dunn, A. L., & Fox Tree, J. E. (2009). A quick, gradient Bilingual Dominance Scale. *Bilingualism: Language and Cognition*, 12(03), 27-

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34*(2), 153-175.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & cognition*, 19(5), 448-58.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2011). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, 1-22.

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.

Kalikow, D., Stevens, K. N., & Elliott, L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5), 1337.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, 60(3), 355-76.

Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2011). A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. *Journal of Phonetics, 39*(2), 156-157. d

Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician enhancement for speech-in-noise. *Ear and hearing*, 30(6), 653-61.

Patel, A. D. (2011). Why would Musical Training Benefit the Neural Encoding of Speech? The OPERA Hypothesis. *Frontiers in psychology*, *2*, 142.

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of Musical Disorders The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Science*, 999, 58-75.

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

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

Slevc, L. R., & Miyake, A. (2006). Individual differences in second language proficiency: Does musical ability matter? *Psychological Science*, *17*(8), 675–681.

Van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: neuroanatomical correlates of phonagnosia. *Journal of clinical and experimental neuropsychology*, 11(5), 665-74.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420-2.