frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 02 September 2022
DOI 10.3389/fpsyg.2022.901326

The own-voice benefit for word recognition in early bilinguals

Sarah Cheung[1] and Molly Babel[2]*

[1]Department of Speech-Language Pathology, University of Toronto, Toronto, ON, Canada,
[2]Department of Linguistics, University of British Columbia, Vancouver, BC, Canada

Check for updates

OPEN ACCESS

EDITED BY
Joseph V. Casillas,
Rutgers, The State University
of New Jersey, United States

REVIEWED BY
Pui Fong Kan,
University of Colorado Boulder,
United States
Linda Polka,
McGill University, Canada

*CORRESPONDENCE
Molly Babel
molly.babel@ubc.ca

SPECIALTY SECTION
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 21 March 2022
ACCEPTED 28 July 2022
PUBLISHED 02 September 2022

CITATION
Cheung S and Babel M (2022) The own-voice benefit for word recognition in early bilinguals.
Front. Psychol. 13:901326.
doi: 10.3389/fpsyg.2022.901326

The current study examines the self-voice benefit in an early bilingual population. Female Cantonese–English bilinguals produced words containing Cantonese contrasts. A subset of these minimal pairs was selected as stimuli for a perception task. Speakers' productions were grouped according to how acoustically contrastive their pronunciation of each minimal pair was and these groupings were used to design personalized experiments for each participant, featuring their own voice and the voices of others' similarly-contrastive tokens. The perception task was a two-alternative forced-choice word identification paradigm in which participants heard isolated Cantonese words, which had undergone synthesis to mask the original talker identity. Listeners were more accurate in recognizing minimal pairs produced in their own (disguised) voice than recognizing the realizations of speakers who maintain similar degrees of phonetic contrast for the same minimal pairs. Generally, individuals with larger phonetic contrasts were also more accurate in word identification for self and other voices overall. These results provide evidence for an own-voice benefit for early bilinguals. These results suggest that the phonetic distributions that undergird phonological contrasts are heavily shaped by one's own phonetic realizations.

KEYWORDS

speech perception, word recognition, bilingualism, speech production, linguistic representation

Introduction

Familiar accents and voices receive a range of processing benefits including higher recognition rates, intelligibility boosts, and increased attention in the context of competing speech (e.g., Bradlow and Bent, 2008; Adank et al., 2009; Johnsrude et al., 2013; Holmes et al., 2018). One's own voice is arguably the most familiar voice, due to our continuous exposure to it. Given that self-recognition, the ability to distinguish between the self and others, is a fundamental human capability, it is therefore unsurprising that self-referential information is processed differently from stimuli associated with others across domains (Keenan et al., 2000; Platek et al., 2004, 2006; Uddin et al., 2005; Keyes et al., 2010; Devue and Brédart, 2011; Zhao et al., 2011; Liu et al., 2019). This extends to voice processing, as researchers have not only observed that people process their own voices differently from others' voices (Hughes and Harrison, 2013; Peng et al., 2019;

Mitterer et al., 2020), but also that this difference in perception may translate into an advantage in recognizing words in self-produced speech (Eger and Reinisch, 2019).

Spoken language processing is, in a large part, shaped by experience. Infants narrow their perceptual categories based on the language varieties they are exposed to (e.g., Werker and Tees, 1984), and adults prioritize phonetic information in a language-specific manner (e.g., Johnson, 1997; Sumner et al., 2014; Schertz and Clare, 2020). Familiar languages, accents, and voices are afforded benefits in processing, and these benefits surface at different intervals in the pipeline. Concepts like *recognition* (i.e., comprehending the signal) and *encoding* (i.e., updating a representation) are different processes (Clopper et al., 2016; Todd et al., 2019) and consideration needs to be given as to whether any socially skewed or preferential encoding takes place at *perception* or *interpretation* stages (see Zheng and Samuel, 2017). In addition to unpacking the mechanisms by which preferential encoding occurs, the acoustic-auditory substance of *what* is preferentially encoded is not well predicted by theory or supported by consistent empirical results. For example, while there is evidence that familiar speech signals are preferentially encoded (e.g., Clopper et al., 2016), this does not entail that the highest frequency exemplar is the most robustly encoded (Sumner and Samuel, 2005). In some cases, early and consistent experiences shape recognition (e.g., Sumner and Samuel, 2009) and perceptual processing (Evans and Iverson, 2007), whereas in other instances, socially prestigious speech may receive a boost (Sumner and Kataoka, 2013). Familiar accents typically receive benefits, but unfamiliar accents can draw perceptual attention, making them more challenging to ignore than more familiar accents (Senior and Babel, 2018).

The aforementioned examples all relate to accent or dialect differences, but familiarity effects in spoken language are not limited to that level of abstraction. Familiarity effects also extend to individual voices. A large body of research demonstrates that familiarity with a speaker's voice eases perception (Nygaard et al., 1994; Newman et al., 2001; Perry et al., 2018). For instance, Nygaard and Pisoni (1998) showed that listeners who successfully learned the voices and names of speakers were better at identifying speech produced by the speakers they were trained on compared to unfamiliar speakers. Evidence of a familiar-talker advantage in perception has been found for young and old listeners (Yonan and Sommers, 2000; Johnsrude et al., 2013), in addition to older listeners with hearing impairments (Souza et al., 2013). Familiar-talker advantages are also found with explicit (Nygaard and Pisoni, 1998) and implicit training (Kreitewolf et al., 2017), as well as in listening conditions with a competing talker in the background (Holmes et al., 2018; Holmes and Johnsrude, 2020). Listeners show improved abilities to selectively attend to or ignore very high familiarity voices (e.g., a spouse's voice; Johnsrude et al., 2013), suggesting that a relatively fine-grained prediction is available for familiar voices. Even without awareness of speaker identity, listeners encode acoustically-specific information about words, which can result in more efficient processing if it is similar to existing representations (Creel and Tumlin, 2011).

As noted, an individual's own voice is, arguably, the voice that one has most familiarity with. Importantly, however, self-voice perception of one's own voice "sounds different" from others' because of the different mediums through which sound is physically conducted during perception. When listeners hear their own voices as they speak, sound is transmitted via both air and bone conduction (Shuster and Durrant, 2003; Reinfeldt et al., 2010). In air conduction, vibrations exit the oral cavity, travel through air and enter the ear canal, whereas in bone conduction, vibrations move through the skull bone directly to the cochlea (Stenfelt and Goode, 2005). Comparatively, when listeners hear others speak or hear their own voice in recordings, sound is conducted solely via air conduction. Despite these differences, listeners are very successful at recognizing their own productions in recordings (Xu et al., 2013). Xu et al. (2013) presented listeners with recordings of their own voices and the voices of other, familiar speakers in normal and difficult listening conditions. They found that even in high-pass filter conditions that removed acoustic information from the mean of an individual's third resonant frequency and above, listeners were able to identify their own voices. Researchers theorize that auditory familiarity with one's own voice and the association between auditory self-representation and motor representations may contribute to this self-recognition advantage (Xu et al., 2013).

Beyond an advantage in own-voice recognition, speakers monitor their own productions through auditory feedback. Delayed auditory feedback induces an increase in foreign accent for second language learners (Howell and Dworzynski, 2001) and an increase in regional accent for those who have acquired a different accent (Howell et al., 2006). This suggests that when the timing of auditory feedback is perturbed, individuals are unable to monitor their speech as effectively, resulting in a shift in their speech patterns. Real-time shifts in auditory feedback, where an individual hears resynthesized versions of their own productions that deviate from what they produced, elicits compensation to account for the synthesized acoustic shift (e.g., Houde and Jordan, 2002; Jones and Munhall, 2002; Purcell and Munhall, 2006; Katseff et al., 2012). Crucially, the magnitude of an individual's compensatory response is associated with the shifted item's position in the vowel space; shifted items that fall near a phonetic category boundary elicit a larger compensatory response (Niziolek and Guenther, 2013). Compensation for auditory feedback appears to be generally heightened for linguistically relevant dimensions (Xu et al., 2004; Chen et al., 2007; Mitsuya et al., 2011; Niziolek and Guenther, 2013).

While one's own auditory feedback is valuable to the control of motor actions in speech, do one's own productions provide a recognition advantage at the word level? Word recognition can be considered a process that serves to comprehend the speech of *others*, as, under normal contexts, an individual is aware of the linguistic message that is emitted from their own vocal tract. We are interested in how own-voice familiarity shapes the representational and recognition space for linguistic contrasts in word recognition and the acoustic-phonetic distributions that implement phonological contrasts. To test how one's own implementation of a contrast affects word recognition, an introduction of some kind of adverse listening condition is required, as identifying words in a familiar language is a fairly trivial task. Scholars have approached this from two angles – with second-language (L2) learners or first language listeners – each of which has used relatively distinct methods and landed on different conclusions.

From the L2 perspective is Eger and Reinisch (2019), who demonstrated that German-speaking learners of English were better at recognizing self-produced words in English. This suggests that L2 language learners prioritize their own realizations of phonological contrast. In a related study, Mitterer et al. (2020) show that German-speaking learners of English rate their own, in this case, vocally disguised, sentence productions as more target-like. Mitterer and colleagues offer the interpretation that it is the comprehension advantage afforded by one's own voice that supports higher ratings for self-produced sentences. However, these results for L2 language learners contrast with claims made when processing a first language. For an individual's first language, there is a reported benefit to processing the most statistically average voice over their own self-produced voice when listeners are asked to identify noise-vocoded words, a manipulation that removes fine spectral detail, but spares temporal cues and amplitude modulation (Schuerman et al., 2015, 2019). There is, however, some evidence that L1 listeners' word recognition in sentences masked with speech-shaped noise shows a benefit for self-produced sentences compared to sentences produced by others (Schuerman, 2017). Schuerman et al. (2015, 2019) suggest that listeners' preferred linguistic representations are informed by the input perceived in one's speech community — hence the improved recognition for the statistically average voice in noise-vocoded speech. They reason that own-voice preferences may only arise when listeners are aware that they are hearing their own voice, which is challenging in noise-vocoded speech. The mechanism for the own-voice benefit for L2 English learners posited by Eger and Reinisch (2019) presumes that an individual recognizes their own voice and then further perceptually adapts to their own productions.

In the current study, we test the own-voice benefit for word recognition in early bilinguals, leveraging the high levels of natural phonetic ambiguity in a heterogenous multilingual population of Cantonese–English speakers. We test whether these early bilinguals, like second language learners, show an own-voice benefit in word recognition. Moreover, we probe whether the own-voice benefit indeed hinges upon recognition of one's own voice. Following prior work (Holmes et al., 2018; Mitterer et al., 2020), some cues to talker identity are manipulated by shifting f0 and formant frequencies (using Praat; Boersma and Weenink, 2020) to limit listeners' ability to recognize their own voices. This methodology draws on the observation that manipulating these cues greatly affects the success of self-voice recognition (Xu et al., 2013).

Materials and methods

The experiment consisted of three parts: a questionnaire about multilingualism, a production task, and a perception task, all of which were completed remotely on participants' own electronic devices. All written and verbal instructions were presented in English to accommodate limited Cantonese literacy within the bilingual population at our university.

Participants

To be eligible for this study, participants were required to self-identify as female, be exposed to both Cantonese and English at or before the age of six, and minimally have the ability to carry out a basic conversation in Cantonese. Only female subjects were invited to participate to minimize between-speaker variation and to allow a more consistent vocal disguise technique (see description of audio manipulation below). Thirty-six female Cantonese-English bilinguals participated in the experiment. While all participants completed the multilingual questionnaire and the production task, the recordings of three participants obtained during the production task were excluded from the perception task due to poor recording quality and interference from background noise. In addition, two participants who completed the production task and questionnaire did not complete the perception task, resulting in 31 subjects who completed all three parts of the study. **Appendix Table A1** provides selected summary language information for the 33 participants who completed the production task and for whom a perception experiment was designed. **Appendix Table A2** contains additional demographic information about the participant population. Participants reported their languages in order of current self-assessed dominance, along with the age of acquisition of each language, and speaking, listening, and reading proficiencies on a scale from 0 (none) to 10 (perfect). The population is highly multilingual, as is typical of both Cantonese speakers in Cantonese-speaking homelands (e.g., Hong Kong, Guangzhou) and those in the Cantonese-speaking diaspora, which is the convenience sample used in the current study. For example, 27 participants report Mandarin as an additional language, and 16 report French, in addition to small numbers of individuals self-reporting knowledge of other languages. Participants' self-reported ages of acquisition indicate that Cantonese was the earliest acquired language (Median = 0, *SD* = 1.3), compared to English (Median = 3, *SD* = 1.9), and Mandarin (Median = 6, *SD* = 4) and French (Median = 9, *SD* = 2.7), the other two most attested languages amongst participants. Participants self-reported significantly higher speaking and listening proficiencies for English (speaking: *M* = 9.3, *SD* = 0.98; listening: *M* = 9.48, *SD* = 0.83) compared to Cantonese [speaking: *M* = 7.15, *SD* = 2.36; listening: *M* = 7.82, *SD* = 1.96; paired *t*-test for speaking: *t*(32) = 4.38, *p* = 0.0001; paired *t*-test for listening: *t*(32) = 4.11, *p* = 0.0003]. Mandarin was the language with the next highest self-reported proficiency across participants, though it was not a language reported by all participants, and self-reported speaking [unpaired *t*-test: *t*(54) = –2.65, *p* = 0.01] and listening [unpaired *t*-test: *t*(54) = –2.7, *p* = 0.009] skills were higher for Cantonese than Mandarin (speaking: *M* = 5.5, *SD* = 2.5; listening: *M* = 6.4, *SD* = 2.1). Participants' current place of residence was in English-dominant communities in Canada and the United States, as shown in **Appendix Table A2**.

Participants were compensated with gift cards equivalent to $5 CAD for the production task, $5 CAD for the questionnaire, and $10 CAD for the perception task. Participants were recruited through the UBC community and social media.

Materials

Multilingual language questionnaire

Participants completed an online survey that presented questions from the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007) and the Bilingual Language Profile (BLP; Gertken et al., 2014). Both resources were designed to gain a better understanding of language profiles of bilingual and multilingual speakers by including questions relating to individuals' language history, usage, attitudes and self-rated proficiency. Additionally, general questions pertaining to participants' biographical information were included in this questionnaire. This survey was administered in English.

Production stimuli

Stimuli for the production task included monosyllabic Cantonese words, presented as pictures accompanied by English translations. All pictures were hand-drawn by the researcher and presented in black and white so that no single picture was especially salient to subjects (see **Appendix Figure A2** for the complete set of visual stimuli). The word list was composed of 22 minimal pairs targeting seven segmental contrasts (see **Appendix Table A3** for the complete production word list). Three of the lexical items served as minimal pair to more than one other item, hence the number of unique words totaled 41 (and not 44) for the 22 minimal pairs. The lexical items involved word initial consonants /t͡s/, /t͡sʰ/, and /s/ and vowel contrasts /ɐ͡i/ and /e͡i/, /ɔː/ and /o͡u/, /ɐ͡i/ and /aː͡i/, /ɐ͡u/ and /aː͡u/, and /ɐ/ and /a:/. Target sounds were selected based on their presence in Cantonese and absence in English such that the selected contrasts would show variability across proficiency ranges in the Cantonese-English bilingual community. For example, three of the vowel contrasts chosen are distinguished by vowel length, a feature that is not lexically contrastive in English. The stimuli were designed to consist of all high level tone (T1) words to control for differences in tone that may cause unwanted variability in production or confusion in perception task performance. The words were chosen to be familiar to Cantonese speakers with potentially limited vocabularies due to largely using Cantonese as a home language in an English-dominant region and had meanings that could be easily represented in pictures. Pictures, as opposed to Chinese characters, were used both in the production and perception tasks to accommodate participants who have limited literacy skills.

Perception stimuli

A subset of the stimuli words used in the production task were featured in the perception task. These consisted of 13 minimal pairs featuring five vowel contrasts: /ɐ͡i/ and /e͡i/, /ɔː/ and /o͡u/, /ɐ͡i / and /aː͡i/, /ɐ͡u/ and /aː͡u/, and /ɐ/ and /a:/, which are presented in their character and Jyutping transliterations and English glosses in **Table 1**. The same pictures corresponding to these target words from the production experiment were used in the perception task. The manipulation of the audio stimuli for the perception experiment is described below.

Production task

Procedure

For the production task, participants first watched a video tutorial (made by the first author) on how to record themselves producing the list of target words. This video included a familiarization phase for participants to learn the intended referents of the picture stimuli. For each target word, participants would hear a Cantonese word and see its corresponding picture and English translation. Afterward, participants were instructed to download Praat (Boersma and Weenink, 2020) and record themselves using the built-in microphone of their personal electronic devices at a sampling frequency of 44,100 Hz. Participants accessed a .pdf file containing the picture stimuli and were asked to verbally label the target words in Cantonese, given the picture and English translation as they proceeded through the randomized list at their own pace. Each picture was shown twice to elicit two productions of each word, for a total of 82 productions. Lastly, participants were asked to verbally describe a picture

TABLE 1 *Perception Stimuli* arranged by minimal pair.

Chinese Character	English Gloss	Jyutping Romanization	Chinese Character	English Gloss	Jyutping Romanization
雞	chicken	gai1	機	machine	gei1
雞	chicken	gai1	街	street	gaai1
揮	to wave	fai1	飛	to fly	fei1
多	many	do1	刀	knife	dou1
歌	song	go1	高	tall	gou1
梳	comb	so1	鬚	beard, moustache	sou1
波	ball	bo1	煲	pot	bou1
踎	to squat	mau1	貓	cat	maau1
秋	autumn	cau1	抄	to copy	caau1
咯	cough	kat1	咭	card	kaat1
心	heart	sam1	衫	shirt	saam1
西	west	sai1	嘥	to waste	saai1
龜	turtle	gwai1	乖	well-behaved	gwaai1

Note that 雞 *chicken* is used in two minimal pairs.

of a busy park scene in Cantonese, in as much detail as they wanted. Participants saved their recordings according to their anonymous participant ID number and uploaded their recordings to Dropbox.

Segmentation

Words of the minimal pairs were segmented from recordings using Praat (Boersma and Weenink, 2020). Recordings from three participants were excluded from this process due to poor recording quality. From the productions of the remaining 33 speakers, nine speakers had at least one word excluded for a total of 15 words excluded from analyses due to incorrect labeling of the picture stimuli. The removal of one item entailed the removal of two, as the minimal pair was removed from that individual's set.

Because stimuli words were produced in isolation, word-initial stops /b/, /d/, /g/, /k/ and /kʷ/ were identified as beginning with the stop burst, starting as an abrupt change in amplitude in the waveform and ending with the onset of quasi-periodic activity of the following vowel. The offset of the labialized voiceless velar stop /kʷ/ was identified as a change in the waveform from a simpler periodic pattern to a more complex periodic pattern of a vowel. In this set of stimuli, the only word-final stop was /t̚/. The end boundary of this unreleased stop was identified as the same point as the end of its preceding vowel. Fricatives /s/ and /f/ were identified in waveforms as aperiodic or random patterns indicating frication noise. Affricates /t͡s/ and /t͡sʰ/ were identified as beginning with a stop burst and ending with the offset of frication noise, signaling the end of the fricative. Aspirated alveolar affricates showed a period of high amplitude frication followed by a period of lower amplitude frication and the boundaries for aspiration were annotated using low amplitude frication as a cue. One participant produced target words intended to contain word-initial aspirated alveolar affricates with voiceless fricatives instead. For these productions, the onset and offset of the aspirated alveolar affricate /t͡sʰ/ were marked at the same points as the beginning and end of aspiration shown in the waveform. The onset of nasals /m/, /n/ and /ŋ/ were identified at the point of a most discrete change in amplitude in the waveform. The offset of the nasal consonants in word-initial position were indicated by a sudden increase in intensity at the beginning of the following vowel. Another cue used to identify this boundary was the change from a simple waveform pattern with lower frequencies, characteristic of nasal consonants, to a more complex pattern with both high and low frequencies, characteristic of vowels. Likewise, the opposite change in intensity and opposite shift in waveform patterns indicated boundary of the word-final nasal /ŋ/. All word and sound boundaries were placed as closely as possible to zero crossings to prevent auditory distortions resulting from discontinuities at the beginnings and ends of sound intervals. Words in all 22 minimal pairs were segmented, although only the subset of words comprising 13 minimal pairs were used in the perception task. Target words were saved into their own files, while target sounds were trimmed into files with 25 ms buffers at the onset and offset of sounds in preparation for acoustic analysis.

Grouping voices

Acoustic analyses served to group minimal pairs into five groups (Groups A, B, C, D, and E) reflecting how discretely speakers produced the contrast between the two words of each minimal pair. We will refer to this measure as "contrastiveness," as it denotes the acoustic difference between target sounds in minimal pairs, but does not necessarily imply speaker proficiency or production accuracy. Because of the considerable amount of individual variation observed between minimal pairs within vowel contrasts, a given talker's group assignment was done separately for each minimal pair. This means that a speaker was not, for example, categorized as a Group A speaker, but her productions for a particular minimal pair may have been assigned to Group A, while her productions for other minimal pairs may be in another contrastiveness Group.

To determine contrastiveness we first estimated formant trajectories with samples every two seconds for each vowel using Fast Track (Barreda, 2021), a formant tracker plug-in via Praat (Boersma and Weenink, 2020). The frequency range was set at 5,000–7,000 Hz to reflect a speaker of "medium height" (Barreda, 2021), as all participants in our study were female adults.

Formant trajectories were then converted from Hertz to the Bark scale to better reflect auditory processing (Traunmüller, 1990). With the obtained Bark-scaled formant trajectories, we then performed a discrete cosine transform (DCT) which yielded three primary coefficients for F1 and F2. The three coefficients corresponded to the mean of the formant, the slope of the formant and the curvature of the slope. In addition to these six dimensions, we also measured vowel duration as a seventh dimension in which speakers could potentially show distinctiveness in production. While not all seven dimensions may be used to contrast the target vowels in our minimal pairs, we did not exclude any particular parameter to avoid making any *a priori* claims about the relative importance of these cues for contrastiveness for this bilingual population. We centered, scaled and calculated Euclidean distances for each talker's minimal pair along all seven dimensions.

Lastly, for each minimal pair, we organized speakers according to the contrastiveness of their productions. This was done by ranking the Euclidean distances for each minimal pair and using the rankings of each to form minimal pair-specific group assignments, in which a greater Euclidean distance indicated a more distinctive production. Within each minimal pair, we formed five groups, ranging from A (most contrastive) to E (least contrastive), consisting of five to seven different voices; thus, for each minimal pair, each group had 5-7 different voices. The groups were manually adjusted to be approximately equally sized, as some talkers were missing tokens and therefore would not be presented with that particular minimal pair in their individualized perception experiment. **Figure 1** is a box-and-whisker plot presenting the phonetic distance or contrastiveness range for the productions in each of the five contrastiveness groups.

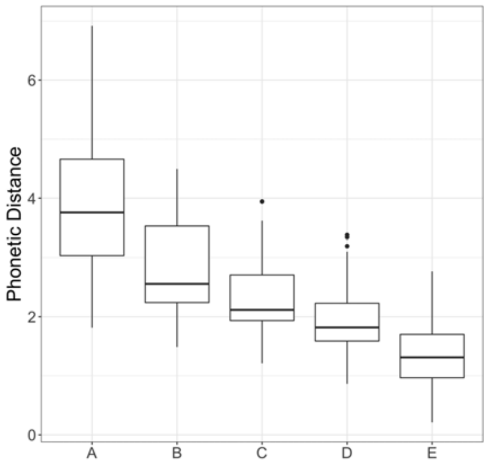

FIGURE 1
Box-and-whisker plot of phonetic distance between minimal pairs for utterances in the five contrastiveness groups.

Each subject was presented with a perception experiment, described below, featuring their own productions and the productions of other members of their contrastiveness group, for each minimal pair. Therefore, the number of different unfamiliar voices heard by each participant varied according to their group memberships.

Perception task

Audio manipulation

For the perception experiment, recordings segmented into isolated words were altered to change female voices into male-like voices using the Change-Gender function in Praat (Boersma and Weenink, 2020). This application lowered the fundamental frequency (f0) and formant frequencies of the original productions by multiplying these dimensions by factors specific to each speaker. Modulation of these parameters have been shown to influence the accuracy of self-voice recognition (Xu et al., 2013) and previous studies have successfully disguised voices using the Change-Gender function (Holmes et al., 2018; Mitterer et al., 2020). For speakers in the current study, the multiplication factors for f0 and formant frequencies ranged from 0.55 to 0.75 (mean = 0.62) and 0.79 to 0.83 (mean = 0.81) respectively. Pitch range parameters were adjusted as necessary to ensure accurate pitch tracking. Following Mitterer et al. (2020), the manipulations started with scaling the f0 by 0.59 and the formants by 0.82, which were the average manipulations made by Mitterer et al. (2020). From there, the actual values for each talker were adjusted by ear to achieve a good-sounding disguise. The specific by-talker adjustments are reported in **Appendix Table A4**. Finally, the target stimuli were RMS-amplitude normalized to 65 dB and mixed in continuous speech-shaped noise, created from the spectral profiles of the participants' speech samples, at a signal-to-noise ratio (SNR) of +5 dB to increase the difficulty of the task. This particular SNR was determined through piloting to achieve high accuracy, but prevent ceiling performance.

Procedure

The same speakers who completed the production task were invited to complete the perception task several months later, which was administered online using jsPsych (de Leeuw, 2015).

This perception experiment was a two-alternative forced-choice lexical identification task featuring the acoustically altered recordings described above. For each trial, participants heard an isolated Cantonese word produced either by themselves or another speaker along with two pictures on the left and right sides of the screen, representing the appropriate Cantonese minimal pair. Participants were required to choose the picture corresponding to the word they heard by pressing the keys "F" or "J" for the left and right sides of the screen, respectively. Participants' responses advanced the program to the next trial. Three practice trials were provided. Audio stimuli were presented at a comfortable listening level and participants completed a headphone check prior to beginning the experiment (Woods et al., 2017). There were four repetitions of each token for a total of 560–688 trials for each participant's personalized experiment [up to 26 items (e.g., 13 minimal pairs) × a range of 5–7 speakers in each by minimal pair group × 4 repetitions of each token]. Trials were fully randomized across four blocks between which participants were offered a self-paced break. At the end of the experiment, participants were asked if they recognized their own voice throughout the experiment, to which they selected "yes" or "no" on the screen. The perception experiment was completed on participants' own electronic devices and took approximately 35–40 min to complete. Participants were asked to complete the task in a quiet place.

Results

To remove extremely fast and extremely slow responses, button presses logged under 200 ms and over 5000 ms were removed from the data, eliminating just under 2% of responses. Participants' responses on the perception task were scored as either correct or incorrect depending on whether listeners chose the picture corresponding to the intended word. These accuracy data were analyzed using a Bayesian multilevel regression model in Stan (Gabry and Cešnovar, 2021) using brms (Bürkner, 2018) in R (R Core Team, 2021). The accuracy of each response (correct word identification or not) was analyzed as the dependent variable with Voice Match (other voice, own voice), Trial number (centered and scaled), and Contrastiveness Group (Groups A–E) as independent variables. Voice Match and Group, Trial and Group, and Trial and Voice Match were included as interactions. There were random slopes for Voice Match and Trial by participant. Given that most items were other voice items, Voice Match was treatment coded (with Other Voice as the reference level) and Contrastiveness Group was forward-difference coded using the coding matrices package (Venables, 2021), which compares each level in Contrastiveness Group to the adjacent level. The model family was Bernoulli and we specified weakly informative normally distributed priors that were centered at 0 for the intercept and population-level parameters. The intercept and population-level parameters had standard deviations of 5 and 2.5, respectively, following recommendations for accuracy data in Coretta (2021). Correlations used the LKJ prior with a value of 2. The models were fit with 4000 iterations (1000 warm-up) with four chains for the Hamiltonian Monte-Carlo sampling. All R-hat values were below 1.01 and Bulk ESS values were all high, suggesting the model was well mixed. The median posterior point estimates and the 95% credible interval (CrI) is reported for all parameters and interactions. An effect is considered compelling if 95% of the posterior distribution for a parameter does not include 0. An effect is considered to have weak evidence if the credible interval includes 0, but the probability of direction is at least 95%. These interpretation practices follow recommendations in Nicenboim and Vasishth (2016).

The model results are reported in **Table 2**. The intercept indicates that listeners were very good at the task, reliably identifying the intended lexical item [β = 1.66, 95% CrI = [1.32, 2.02], Pr(β > 0) = 1]. The model results provide compelling evidence for a benefit in processing one's own (disguised) voice [β = 0.23, 95% CrI = [0.06, 0.42, Pr(β > 0) = 99.5%]. This result is visualized in **Figure 2**, which presents the fitted draws from the posterior fit of the model for the own-voice effect by Contrastiveness Group.

An effect of trial suggests that listeners' accuracy improved across the course of the experiment [β = 0.07, 95% CrI = [0.01, 0.14], Pr(β > 0) = 98.22%]; the CrI for all interactions of Trial with the Contrastiveness Group contrasts overlap substantially with 0, suggesting that this cross-experiment improvement was not specific to a particular Group. The Voice Match by

TABLE 2 Summary of the posterior distribution modeling word recognition accuracy with posterior means and the 95% Credible Interval, along with the probability of direction for each effect.

Parameter	β	95% CrI	Probability of direction
Intercept	1.66	[1.32, 2.02]	100%
Voice Match (Own Voice)	0.23	[0.06, 0.42]	99.5%
Trial	0.07	[0.01, 0.14]	98.22%
Group A vs. B	−0.21	[−0.36, −0.06]	99.72%
Group B vs. C	0.27	[0.13, 0.41]	100%
Group C vs. D	0.21	[0.07, 0.34]	99.84%
Group D vs. E	0.26	[0.13, 0.39]	100%
Voice Match × Group A vs. B	0.31	[−0.09, 0.70]	93.69%
Voice Match × Group B vs. C	−0.41	[−0.77, −0.04]	98.60%
Voice Match × Group C vs. D	0.31	[−0.04, 0.68]	95.55%
Voice Match × Group D vs. E	−0.04	[−0.37, 0.28]	60.03%
Trial × Group A vs. B	0.08	[−0.06, 0.22]	86.60%
Trial × Group B vs. C	−0.04	[−0.17, 0.09]	73.05%
Trial × Group C vs. D	−0.03	[−0.15, 0.09]	68.46%
Trial × Group D vs. E	−0.04	[−0.15, 0.08]	73.08%
Voice Match × Trial	0.03	[−0.10, 0.17]	65.33%

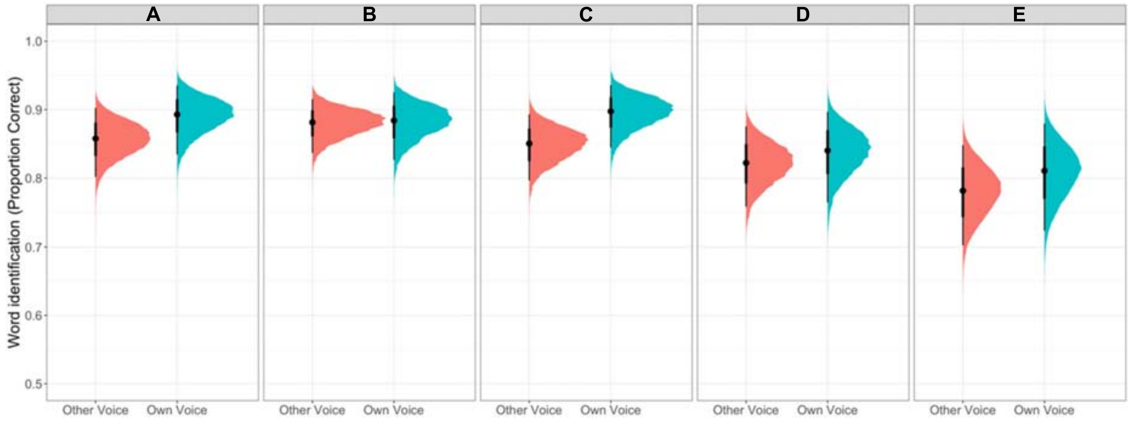

FIGURE 2
Proportion of correct responses in the perception task for the five acoustic contrastiveness groups presented as fitted draws from the posterior fit of the model. Panels **A–E** represent the five contrastiveness groups from most contrastive **(A)** to least contrastive **(E)**. Responses to both own voice and other voices are included.

Trial interaction also overlapped with 0, indicating there is no evidence that the improvement in word recognition across the course of experiment was better or worse for one's own voice or other voices.

Comparisons of adjacent Contrastiveness Groups generally present compelling evidence that higher proficiency groups perform more accurately on the word identification task [Group B vs. C: β = 0.27, 95% CrI = [0.13, 0.41], Pr(β > 0) = 100%; Group C vs. D: β = 0.21, 95% CrI = [0.07, 0.34]; Pr(β > 0) = 99.84%; Group D vs. E: β = 0.26, 95% CrI = [0.13, 0.39], Pr(β > 0) = 100%] with the exception of Group B outperforming Group A [β = –0.21, 95% CrI = [–0.36, –0.06], Pr(β < 0) = 99.72%]. Two interactions involving Voice Match and Group merit attention. There is compelling evidence for an effect that Group B showed less of an own-group advantage than Group C [β = –0.41, 95% CrI = [–0.77, –0.04], Pr(β < 0) = 98.60 %] and there is weak evidence that Group D showed less of an effect than Group C [β = 0.31, 95% CrI = [–0.04, 0.68], Pr(β > 0) = 95.6%].

Discussion

This experiment tested an own-voice advantage for word recognition in Cantonese for Cantonese-English early bilinguals. Words were presented in speech-shaped noise at +5 dB SNR to make the task challenging enough to inhibit ceiling performance. Listeners were more accurate at identifying difficult vowel contrasts if they were (vocally disguised) self-produced items compared to items produced by other individuals who manifested the phonological contrast to a similar degree. This was true despite an individual's own voice being disguised, suggesting that the own-voice word recognition benefit leverages linguistic representations that exist in a normalized representational space, as opposed to relying on an exact acoustic-auditory match to one's natural acoustic patterns. Items were organized by the degree of phonetic distance for the phonological contrast into what are labeled contrastiveness groups. There was strong evidence that Group C showed more of an own-voice benefit than Group B and weak evidence that Group C showed a greater own-voice benefit than Group D. Group B was exceptional in stepping out of the anticipated order in overall accuracy. While it was generally the case that groups with higher contrastiveness performed more accurately on the word identification task, Group B out-performed Group A, the highest contrastiveness group. A possibility for why those in Group B were so outstanding may relate to imperfection in our method of calculating acoustic distance, which included acoustic dimensions that are likely not core cues to contrast, though this is speculation. We note that the overall pattern was that the own-voice benefit was robust across contrastiveness groups and word recognition accuracy decreased as contrastiveness was reduced. The contrastiveness groups relate to the degree of distinctiveness of speakers' productions, which in turn may relate to speaker proficiency. Like the finding in Eger and Reinisch (2019), however, the own-voice benefit does not seem to hinge on proficiency.

Word recognition accuracy improved over the course of the experiment with participants' own voices and other voices. Although subjects heard their own voice more often than any single other voice in the experiment, the proportion of correct responses increased across trials for all voices. Altogether, this suggests that the observed self-advantage was not simply due to listeners hearing their own voice more than other voices throughout the task. The improvement across the experiment was likely due to participants adapting to the noise, which masked the speech to inhibit ceiling performance.

Our ability to determine whether listeners explicitly heard their own voice was based on an explicit self-assessment. A subset of participants reported hearing their own voice in the experiment (n = 9), but we cannot (a) confirm that positive responses to this question were not a function of positive response bias or (b) rule out that other listeners did not implicitly hear their own voices. While we follow previous work in our implementation of the voice disguise (Holmes et al., 2018; Mitterer et al., 2020), an individual's voice identity is available in other spectral and temporal patterns. Speakers vary in terms of their unique voice profiles (Lee et al., 2019; Johnson et al., 2020) and listeners exploit different acoustic cues for talker identification (Van Lancker et al., 1985; Lavner et al., 2000). Schuerman et al. (2015, 2019) did not find support for an own-voice advantage within an individual's first language when presenting noise-vocoded speech, a type of degradation in which many spectral cues important to talker identification are severely reduced, though Schuerman (2017) finds some evidence for an own-voice benefit for word recognition in sentences for speech in noise, which better retains talker-specific information. The removal of expected cues to speaker identity does not explain the absence of an own voice-benefit in those studies, however, as voice recognition and speech recognition are separate, but connected systems (for an overview see Creel and Bregman, 2011). Listeners show an intelligibility benefit for familiar voices even when those voices are made unfamiliar, indicating that the familiarity benefit does not rely on explicit recognition of a voice (Holmes et al., 2018).

The prevalent theory in voice representation is that talkers' voices are represented according to prototypes. According to the prototype theory, each stimulus is compared to a representative or central member of its category; stimuli that better approximate the prototype will be more easily perceived as belonging to the category (Lavner et al., 2001). Under this interpretation, talker identification relies on the storage and retrieval of identities based on a set of features deviating from the prototype. As previous studies have shown, the acoustic dimensions used to characterize different voices are often talker-specific (Van Lancker et al., 1985; Lavner et al., 2000). Voices that deviate more from the prototype are perceived as more distinct and thus, the more distant a speaker's acoustic features are from the central model, the easier the speaker is to be identified (Lavner et al., 2001; Latinus et al., 2013). This may partially explain the variance in participants' self-reports of hearing their own voices in the current study despite our attempt to disguise vocal identity. Those who successfully identified themselves may have had voices that deviated more from the average template and were therefore easier to recognize. Researchers have proposed that the prototype is an average, commonly encountered, yet attractive voice (Lavner et al., 2001; Latinus et al., 2013; Lavan et al., 2019). Accordingly, this voice should be representative of the listeners' language input and environment, and people of the same linguistic community would be expected to share a similar template (Lavner et al., 2001). The implications for having a voice that approximates listeners' community prototypes with regards to a benefit in word recognition needs to be explored further. In Schuerman et al. (2015, 2019) studies, researchers identified a statistically average speaker among the subjects in their studies to represent the average of the linguistic community. When presented with noise-vocoded speech, native Dutch listeners in their studies showed better recognition of words produced by the statistically average speaker in their sample than the listeners themselves. This implies that the benefit of a prototypical voice may extend beyond the benefit of hearing one's own voice for word recognition.

The core finding in the current work is that listeners were more accurate in recognizing minimal pairs produced in their own (disguised) voice than recognizing the realizations of other speakers who maintain similar degrees of phonetic contrast for the same minimal pairs. These findings with Cantonese-English bilinguals, a population which was targeted to leverage the heterogeneity in pronunciation variation within a native speaker population, replicating and extending the findings for second language learners (Eger and Reinisch, 2019). We present evidence of an own-voice benefit for work recognition, like Eger and Reinisch, but this benefit is seen when voices were disguised and the majority of individuals did not report consciously recognizing their own masked voice.

Crucially, the own-voice advantage in word recognition suggests that the phonetic distributions that undergird phonological contrasts are heavily shaped by one's own phonetic realizations, extending the importance of self-produced items beyond real-time self-monitoring (e.g., Howell et al., 2006; Niziolek and Guenther, 2013). Online compensation for altered auditory feedback indicates that auditory self-monitoring leads to immediate, though incomplete, adjustments in speech production. Importantly, the magnitude of these adjustments is yoked to whether the auditory feedback suggests a linguistic contrast is threatened (Niziolek and Guenther, 2013). This suggests a coupled relationship between perception and production where an individual's representational space for perception and recognition align with the distributional pool available for that individual in production. Many frameworks posit some degree of connection between perception and production with theoretical models differing in terms of how parsimonious perception and production repertoires are, amongst other theoretical differences related to the actual representational space (e.g., Liberman and Mattingly, 1985; Fowler, 1996; Johnson, 1997; Goldinger and Azuma, 2004). Certainly, listeners' abilities to perceive phonetic detail is connected to their abilities to produce contrasts (e.g., Werker and Tees, 1984), but does not wholly limit it (e.g., Schouten et al., 2003). Listeners are well attuned to the distribution of phonetic variation within their speech communities, particularly when that phonetic variation has social value (e.g., Johnson et al., 1999; Hay et al., 2006; Munson et al., 2006; Szakay et al., 2016). A fully isomorphic production and perception system fails to account for how listeners adapt to novel input from other speakers without concomitantly changing their own productions (Kraljic et al., 2008). If perception and production exclusively relied on perfectly mapped mental representations, the reorganization of phonetic space or changes in the weighting of acoustic cues due to perceptual learning should also be observed in that individual's productions, but this is not well supported in the existing literature (Schertz and Clare, 2020).

What mechanism accounts for the own-voice benefit? One possibility is that the mere constant auditory exposure to one's own voice, despite the fact that an individual need not attend to their own speech for the purpose of comprehension, bestows such a high level of familiarity that it is privileged in recognition space. Alternatively, it is plausible that the way in which an individual produces a contrast is intimately tied to the way in which the contrast is realized by their most frequent interlocutors such that this manifestation of the contrast — realized by the most familiar voices and one's own — receives a recognition benefit. This explanation seems unlikely, however, given that second language learners (Eger and Reinisch, 2019) and our early bilingual population show the same own-voice benefit. A third possibility is that while, as described above, perception and production cannot be isomorphic, the yoking of an individual's speech production repertoire and that repertoire's mapping in the perceptual space is what benefits an individual's own-voice productions in recognition. This is also an interpretation offered for own-voice recognition by Xu et al. (2013), who suggest that own-voice auditory and motor representations are connected. The representation of perception and action in shared space is at the heart of the common coding hypothesis (Prinz, 1997). Assuming a shared representational space for perception and production, the common coding theory predicts that listeners compare incoming speech signals to their own productions. Therefore, in perceiving one's own voice, recognition is facilitated because the auditory signal aligns with the listeners' own productions to a greater degree. Support for this in the recognition space comes from speech-reading. Individuals are better at keyword recognition in sentences when speech-reading silent videos of themselves compared to others (Tye-Murray et al., 2013), in addition to receiving more of an audio-visual boost in noisy conditions with their own videos (Tye-Murray et al., 2015). If a shared representational space accounts for the own-voice benefit, it apparently must be part of a developmental trajectory, however, as Cooper et al. (2018) find no evidence for an own-voice benefit (or an own-mother voice) benefit for word recognition in 2.5 year olds (see also Hazan and Markham, 2004). Toddlers are better at recognizing any adult production (their own mother or a different mother) than recognizing self-produced words or words from another toddler. Infants, however, already use sensorimotor information in speech perception. English-acquiring six-month olds' abilities to perceive retroflex and dental stop contrasts is inhibited when a soother blocks tongue movement [Bruderer et al., 2015; see also Choi et al. (2019) for more evidence about the connection between sensorimotor and perceptual processing in infants]. These sets of results suggest that phonemic perception and word-level recognition have different developmental trajectories with respect to the integration of motor and auditory/acoustic information streams. Ultimately, the current study cannot adjudicate between these explanatory mechanisms, but rather provides additional evidence of an own-voice benefit in adult word recognition (Tye-Murray et al., 2013, 2015; Eger and Reinisch, 2019). Multiple threads in the literature do seem to suggest that the integration of production and perceptual representations offers promise in terms of explanatory force.

The proposed mechanism that supports an own-voice benefit in word recognition — the integration of motor and acoustic-auditory representations in the linguistic representations used for word recognition — is not intended to be unique to L2 speech processing (e.g., Eger and Reinisch, 2019) or the processing of one's less dominant language (e.g., the current work). It may simply be easier to observe the evidence of an own-voice benefit in individuals' non-dominant language(s) because it may be more error prone. Individuals' native or dominant languages also, of course, exhibit within- and cross-talker variability (e.g., Newman et al., 2001; Vaughn et al., 2019). It is important to note that while there is strong statistical evidence in support of an own-voice benefit in the current work, the effect is small. An own-voice benefit is also not mutually exclusive with a benefit for a typical voice that represents the prototype or central tendency of the local speech community (e.g., Schuerman, 2017). While listeners are highly adaptable, leveraging any available information in the signal to recognize words, it is important that work in this area use spectrally rich speech samples, as some adverse listening conditions, like noise-vocoded speech, do not encode the full array of spectral information listeners typically have access to in spoken language processing. A degraded signal may encourage listeners to engage in different processing strategies.

While the own-voice benefit for word recognition was statistically robust, some participants did appear to perform less accurately on their own voices. If some aspects of word recognition are related to community averages or prototypes, these individual differences could be accounted for by considering how distant a particular individual is from the prototype. For example, participants exemplifying a self-benefit may better approximate the prototype, while those performing worse with their own voices may deviate more from the prototype relative to other speakers in their group. This reasoning aligns with the Schuerman et al. (2015, 2019) explanation for the benefit bestowed by the statistically average voice. A shared representation for an average speaker in a heterogenous bilingual population presents a challenge, however. In multilingual speech communities where individuals vary in proficiency and language use patterns, which voices are used to form prototypes for which languages? That is, are there separate prototypes, for example, for apparent native speakers of Cantonese and apparent native speakers of English, with separate prototypes established for individuals whose voices suggest a variety of Cantonese-accented English or English-accented Cantonese? What is the representational space for a speaker who experiences speaking and listening to all of these codes in different contexts? We note the nebulous nature of this space, not to discount its importance, but rather to encourage further research that can tackle the complexities in phonetic variation that are experienced by multilingual individuals.

Our recruitment criteria specified exposure to Cantonese from an early age, at or prior to age six. This lumps very early and early acquisition and both simultaneous and sequential bilinguals all in a single group. This may ultimately not be a uniform population. Exposure to a language from birth has implications for pronunciation patterns. For example, Amengual (2019) examined the lenition rates of phrase-initial voice stops and approximants in the Spanish of simultaneous Spanish-English bilinguals, early sequential Spanish-then-English bilinguals, and late Spanish learners (with English as a first language). The simultaneous bilinguals and late learners patterned together. Given that exposure to English from birth unifies these two groups, these results suggest that early exposure to English has the potential to shape pronunciation patterns in adulthood, similar to previous suggestions for perception (e.g., Sebastián-Gallés et al., 2005). The developmental trajectory out of the sensitive period, however, is gradual, and what exactly is the appropriate age delimiter for a particular linguistic representation, pattern, or process is yet to be determined (see, for example, Flege, 1999; Werker and Tees, 2005; Cargnelutti et al., 2019).

Conclusion

Early Cantonese–English bilinguals exhibited an own-voice benefit for word recognition in Cantonese even when self-recognition of their own voice was masked by a vocal disguise. These results complement the evidence indicating an own-voice benefit in second language speakers (Eger and Reinisch, 2019). The own-voice benefit despite overt recognition of one's own voice suggests a coupled relationship between the motor representations and the multidimensional acoustic-auditory representations that support word recognition.

Data availability statement

The listener data supporting the conclusions can be made available by the authors upon request, as that is what aligns with the approved Ethics protocol.

Ethics statement

The studies involving human participants were reviewed and approved by University of British Columbia's Behavioural Research Ethics Committee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SC and MB contributed to the conception and design of the study. SC prepared the stimuli and wrote the first draft. MB performed the statistical analyses and wrote sections of the manuscript. Both authors approved the submitted version.

Funding

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC; Funding reference number AWD-002492).

Acknowledgments

Thanks to Martin Oberg for programming assistance; participants for their time; feedback from all of the members in the Speech in Context Lab, particularly Rachel Soo, Fion Fung, and Roger Lo.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.901326/full#supplementary-material

References

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol.* 35, 520–529. doi: 10.1037/a0013552

Amengual, M. (2019). Type of early bilingualism and its effect on the acoustic realization of allophonic variants: Early sequential and simultaneous bilinguals. *Int. J. Bilingual.* 23, 954–970. doi: 10.1177/1367006917741364

Barreda, S. (2021). Fast track: Fast (nearly) automatic formant-tracking using Praat. *Linguist. Vanguard* 7, 1379–1393. doi: 10.1515/lingvan-2020-0051

Boersma, P., and Weenink, D. (2020). *Praat: Doing phonetics by computer (6.1.21)*. Available online at: http://www.praat.org/ (accessed September 1, 2020).

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13531–13536. doi: 10.1073/pnas.1508631112

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R J.* 10, 395–411. doi: 10.32614/RJ-2018-017

Cargnelutti, E., Tomasino, B., and Fabbro, F. (2019). Language brain representation in bilinguals with different age of appropriation and proficiency of the second language: A meta-analysis of functional imaging studies. *Front. Hum. Neurosci.* 13:154. doi: 10.3389/fnhum.2019.00154

Chen, S. H., Liu, H., Xu, Y., and Larson, C. R. (2007). Voice F0 responses to pitch-shiftedvoice feedback during English speech. *J. Acoust. Soc. Am.* 121, 1157–1163. doi: 10.1121/1.2404624

Choi, D., Bruderer, A. G., and Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of Bruderer et al.(2015). *Psychonom. Bull. Rev.* 26, 1388–1399. doi: 10.3758/s13423-019-01601-0

Clopper, C. G., Tamati, T. N., and Pierrehumbert, J. B. (2016). Variation in the strength of lexical encoding across dialects. *J. Phonet.* 58, 87–103. doi: 10.1016/j.wocn.2016.06.002

Cooper, A., Fecher, N., and Johnson, E. K. (2018). Toddlers' comprehension of adult and child talkers: Adult targets versus vocal tract similarity. *Cognition* 173, 16–20. doi: 10.1016/j.cognition.2017.12.013

Coretta, S. (2021). *Github repository*. Available online at: https://github.com/stefanocoretta/bayes-regression (accessed July 15, 2021).

Creel, S. C., and Bregman, M. R. (2011). How talker identity relates to language processing. *Linguist. Lang. Comp.* 5, 190–204. doi: 10.1111/j.1749-818X.2011.00276.x

Creel, S. C., and Tumlin, M. A. (2011). On-line acoustic and semantic interpretation of talker information. *J. Mem. Lang.* 65, 264–285. doi: 10.1016/j.jml.2011.06.005

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behav. Res. Methods* 47, 1–12. doi: 10.3758/s13428-014-0458-y

Devue, C., and Brédart, S. (2011). The neural correlates of visual self-recognition. *Conscious. Cogn.* 20, 40–51. doi: 10.1016/j.concog.2010.09.007

Eger, N. A., and Reinisch, E. (2019). The impact of one's own voice and production skills on word recognition in a second language. *J. Exp. Psychol.* 45, 552–571. doi: 10.1037/xlm0000599

Evans, B. G., and Iverson, P. (2007). Plasticity in vowel perception and production: A study of accent change in young adults. *J. Acoust. Soc. Am.* 121, 3814–3826. doi: 10.1121/1.2722209

Flege, J. E. (1999). "Age of learning and second language speech," in *Second language acquisition and the critical period hypothesis*, ed. D. Birdsong (London: Routledge), 111–142. doi: 10.4324/9781410601667-10

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *J. Acoust. Soc. Am.* 99, 1730–1741. doi: 10.1121/1.415237

Gabry, J., and Češnovar, R. (2021). *cmdstanr: R interface to 'CmdStan'*. Available online at: https://mc-stan.org/cmdstanr (accessed March 1, 2022).

Gertken, L. M., Amengual, M., and Birdsong, D. (2014). "Assessing language dominance with the bilingual language profile," in *Measuring L2 proficiency: Perspectives from SLA*, eds P. Leclercq, A. Edmonds, and H. Hilton (Bristol: Multilingual Matters), 208–225. doi: 10.21832/9781783092291-014

Goldinger, S. D., and Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonom. Bull. Rev.* 11, 716–722. doi: 10.3758/BF03196625

Hay, J., Warren, P., and Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *J. Phonet.* 34, 458–484. doi: 10.1016/j.wocn.2005.10.001

Hazan, V., and Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *J. Acoust. Soc. Am.* 116, 3108–3118. doi: 10.1121/1.1806826

Holmes, E., and Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *J. Exp. Psychol.* 46, 1465–1476. doi: 10.1037/xlm0000823

Holmes, E., Domingo, Y., and Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychol. Sci.* 29, 1575–1583. doi: 10.1177/0956797618779083

Houde, J. F., and Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *J. Speech Lang. Hear. Res.* 45, 295–310. doi: 10.1044/1092-4388(2002/023)

Howell, P., and Dworzynski, K. (2001). Strength of German accent under altered auditory feedback. *Percept. Psychophys.* 63, 501–513. doi: 10.3758/bf03194416

Howell, P., Barry, W., and Vinson, D. (2006). Strength of British English accents in altered listening conditions. *Percept. Psychophys.* 68, 139–153. doi: 10.3758/bf03193664

Hughes, S. M., and Harrison, M. A. (2013). I like my voice better: Self-enhancement bias in perceptions of voice attractiveness. *Perception* 42, 941–949. doi: 10.1068/p7526

Johnson, K. (1997). "Speech perception without speaker normalization: An exemplar model," in *Talker Variability in Speech Processing*, eds K. Johnson and J. W. Mullennix (San Diego, CA: Academic Press), 145–165.

Johnson, K. A., Babel, M., and Fuhrman, R. A. (2020). *Bilingual acoustic voice variation is similarly structured across languages. Proceedings of Interspeech*. Available online at: https://www.isca-speech.org/archive_v0/Interspeech_2020/pdfs/3095.pdf doi: 10.21437/Interspeech.2020-3095 (accessed October 1, 2020).

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *J. Phonet.* 27, 359–384. doi: 10.1006/jpho.1999.0100

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., and Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychol. Sci.* 24, 1995–2004. doi: 10.1177/0956797613482467

Jones, J. A., and Munhall, K. G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. *J. Phonet.* 30, 303–320. doi: 10.1006/jpho.2001.0160

Katseff, S., Houde, J., and Johnson, K. (2012). Partial compensation for altered auditory feedback: A trade-off with somatosensory feedback? *Lang. Speech* 55, 295–308. doi: 10.1177/0023830911417802

Keenan, J. P., Ganis, G., Freund, S., and Pascual-Leone, A. (2000). Self-face identification is increased with left hand responses. *Lateral. Asymmetr. Body Brain Cogn.* 5, 259–268. doi: 10.1080/713754382

Keyes, H., Brady, N., Reilly, R. B., and Foxe, J. J. (2010). My face or yours? Event-related potential correlates of self-face processing. *Brain Cogn.* 72, 244–254. doi: 10.1016/j.bandc.2009.09.006

Kraljic, T., Brennan, S. E., and Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition* 107, 54–81. doi: 10.1016/j.cognition.2007.07.013

Kreitewolf, J., Mathias, S. R., and von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Front. Psychol.* 8:1584. doi: 10.3389/fpsyg.2017.01584

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., and Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Curr. Biol.* 23, 1075–1080. doi: 10.1016/j.cub.2013.04.055

Lavan, N., Knight, S., and McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nat. Commun.* 10:2404. doi: 10.1038/s41467-019-10295-w

Lavner, Y., Gath, I., and Rosenhouse, J. (2000). Effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Commun.* 30, 9–26. doi: 10.1016/S0167-6393(99)00028-X

Lavner, Y., Rosenhouse, J., and Gath, I. (2001). The prototype model in speaker identification by human listeners. *Int. J. Speech Technol.* 4, 63–74. doi: 10.1023/A:1009656816383

Lee, Y., Keating, P., and Kreiman, J. (2019). Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am.* 146, 1568–1579.

Liberman, A. M., and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition* 21, 1–36.

Liu, L., Li, W., Li, J., Lou, L., and Chen, J. (2019). Temporal features of psychological and physical self-representation: An ERP study. *Front. Psychol.* 10:785. doi: 10.3389/fpsyg.2019.00785

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Mitsuya, T., Macdonald, E. N., Purcell, D. W., and Munhall, K. G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *J. Acoust. Soc. Am.* 130, 2978–2986. doi: 10.1121/1.3643826

Mitterer, H., Eger, N. A., and Reinisch, E. (2020). My English sounds better than yours: Second-language learners perceive their own accent as better than that of their peers. *PLoS One* 15:e0227643. doi: 10.1371/journal.pone.0227643

Munson, B., Jefferson, S. V., and McDonald, E. C. (2006). The influence of perceived sexual orientation on fricative identification. *J. Acoust. Soc. Am.* 119, 2427–2437. doi: 10.1121/1.2173521

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *J. Acoust. Soc. Am.* 109, 1181–1196. doi: 10.1121/1.1348009

Nicenboim, B., and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—Part II. *Lang. Linguist. Comp.* 10, 591–613.

Niziolek, C. A., and Guenther, F. H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *J. Neurosci.* 33, 12090–12098. doi: 10.1523/JNEUROSCI.1008-13.2013

Nygaard, L. C., and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Percept. Psychophys.* 60, 355–376. doi: 10.3758/BF03206860

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychol. Sci.* 5, 42–46. doi: 10.1111/j.1467-9280.1994.tb00612.x

Peng, Z., Wang, Y., Meng, L., Liu, H., and Hu, Z. (2019). One's own and similar voices are more attractive than other voices. *Austral. J. Psychol.* 71, 212–222. doi: 10.1111/ajpy.12235

Perry, L. K., Mech, E. N., MacDonald, M. C., and Seidenberg, M. S. (2018). Influences of speech familiarity on immediate perception and final comprehension. *Psychonom. Bull. Rev.* 25, 431–439. doi: 10.3758/s13423-017-1297-5

Platek, S. M., Keenan, J. P., Gallup, G. G., and Mohamed, F. B. (2004). Where am I? The neurological correlates of self and other. *Cogn. Brain Res.* 19, 114–122. doi: 10.1016/j.cogbrainres.2003.11.014

Platek, S. M., Loughead, J. W., Gur, R. C., Busch, S., Ruparel, K., Phend, N., et al. (2006). Neural substrates for functionally discriminating self-face from personally familiar faces. *Hum. Brain Mapp.* 27, 91–98. doi: 10.1002/hbm.20168

Prinz, W. (1997). Perception and action planning. *Eur. J. Cogn. Psychol.* 9, 129–154.

Purcell, D. W., and Munhall, K. G. (2006). Compensation following real-time manipulation of formants in isolated vowels. *J. Acoust. Soc. Am.* 119, 2288–2297. doi: 10.1121/1.2173514

R Core Team (2021). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Reinfeldt, S., Östli, P., Håkansson, B., and Stenfelt, S. (2010). Hearing one's own voice during phoneme vocalization—Transmission by air and bone conduction. *J. Acoust. Soc. Am.* 128, 751–762. doi: 10.1121/1.3458855

Schertz, J., and Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdiscipl. Rev. Cogn. Sci.* 11, 1–24. doi: 10.1002/wcs.1521

Schouten, B., Gerrits, E., and Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Commun.* 41, 71–80.

Schuerman, W. L. (2017). *Sensorimotor experience in speech perception.* Ph.D. thesis. Nijmegen: Radboud University Nijmegen.

Schuerman, W. L., Meyer, A., and McQueen, J. M. (2015). Do we perceive others better than ourselves? A perceptual benefit for noise-vocoded speech produced by an average speaker. *PLoS One* 10:e0129731. doi: 10.1371/journal.pone.0129731

Schuerman, W., McQueen, J. M., and Meyer, A. (2019). "Speaker statistical averageness modulates word recognition in adverse listening conditions," in *Proceedings of the 19th international congress of phonetic sciences (ICPhS 2019)*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Canberra, NSW: Australasian Speech Science and Technology Association Inc), 1203–1207.

Sebastián-Gallés, N., Echeverría, S., and Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *J. Mem. Lang.* 52, 240–255. doi: 10.1162/jocn.2008.20004

Senior, B., and Babel, M. (2018). The role of unfamiliar accents in competing speech. *J. Acoust. Soc. Am.* 143, 931–942.

Shuster, L. I., and Durrant, J. D. (2003). Toward a better understanding of the perception of self-produced speech. *J. Commun. Disord.* 36, 1–11. doi: 10.1016/S0021-9924(02)00132-6

Souza, P., Gehani, N., Wright, R., and McCloy, D. (2013). The advantage of knowing the talker. *J. Am. Acad. Audiol.* 24, 689–700. doi: 10.3766/jaaa.24.8.6

Stenfelt, S., and Goode, R. L. (2005). Bone-conducted sound: Physiological and clinical aspects. *Otol. Neurotol.* 26, 1245–1261. doi: 10.1097/01.mao.0000187236.10842.d5

Sumner, M., and Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encoding. *J. Acoust. Soc. Am.* 134, EL485–EL491. doi: 10.1121/1.4826151

Sumner, M., and Samuel, A. G. (2005). Perception and representation of regular variation: The case of final/t. *J. Mem. Lang.* 52, 322–338.

Sumner, M., and Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *J. Mem. Lang.* 60, 487–501. doi: 10.1097/WNR.0b013e3283263000

Sumner, M., Kim, S. K., King, E., and McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Front. Psychol.* 4:1015. doi: 10.3389/fpsyg.2013.01015

Szakay, A., Babel, M., and King, J. (2016). Social categories are shared across bilinguals× lexicons. *J. Phonet.* 59, 92–109.

Todd, S., Pierrehumbert, J. B., and Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition* 185, 1–20. doi: 10.1016/j.cognition.2019.01.004

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *J. Acoust. Soc. Am.* 88, 97–100. doi: 10.1121/1.399849

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2013). Reading your own lips: Common-coding theory and visual speech perception. *Psychonom. Bull. Rev.* 20, 115–119. doi: 10.3758/s13423-012-0328-5

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., and Sommers, M. S. (2015). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. *Psychonom. Bull. Rev.* 22, 1048–1053. doi: 10.3758/s13423-014-0774-3

Uddin, L. Q., Kaplan, J. T., Molnar-Szakacs, I., Zaidel, E., and Iacoboni, M. (2005). Self-face recognition activates a frontoparietal "mirror" network in the right hemisphere: An event-related fMRI study. *Neuroimage* 25, 926–935. doi: 10.1016/j.neuroimage.2004.12.018

Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters Part I: Recognition of backward voices. *J. Phonet.* 13, 19–38. doi: 10.1016/s0095-4470(19)30723-5

Vaughn, C., Baese-Berk, M., and Idemaru, K. (2019). Re-examining phonetic variability in native and non-native speech. *Phonetica* 76, 327–358.

Venables, B. (2021). *codingMatrices: Alternative factor coding matrices for linear model formulae. R package version 0.3.3.* Available online at: https://CRAN.R-project.org/package=codingMatrices (accessed March 1, 2022).

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

Werker, J. F., and Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Dev. Psychobiol.* 46, 233–251. doi: 10.1002/dev.20060

Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attent. Percept. Psychophys.* 79, 2064–2072. doi: 10.3758/s13414-017-1361-2

Xu, M., Homae, F., Hashimoto, R., and Hagiwara, H. (2013). Acoustic cues for the recognition of self-voice and other-voice. *Front. Psychol.* 4:735. doi: 10.3389/fpsyg.2013.00735

Xu, Y., Larson, C. R., Bauer, J. J., and Hain, T. C. (2004). Compensation for pitch-shifted auditory feedback during the production of Mandarin tone sequences. *J. Acoust. Soc. Am.* 116, 1168–1178. doi: 10.1121/1.1763952

Yonan, C. A., and Sommers, M. S. (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychol. Aging* 15, 88–99.

Zhao, K., Wu, Q., Zimmer, H. D., and Fu, X. (2011). Electrophysiological correlates of visually processing subject's own name. *Neurosci. Lett.* 491, 143–147. doi: 10.1016/j.neulet.2011.01.025

Zheng, Y., and Samuel, A. G. (2017). Does seeing an Asian face make speech sound more accented? *Attent. Percept. Psychophys.* 79, 1841–1859. doi: 10.3758/s13414-017-1329-2