Bilingualism: Language and Cognition

cambridge.org/bil

Voice processing ability predicts second-language phoneme learning in early bilingual adults

Gaël Cordero[1], Jazmin R. Paredes-Paredes[1], Manuel Perea[2,3], Nuria Sebastian-Galles[4] and Begoña Díaz[1]

[1]Department of Psychology, Faculty of Medicine and Health Sciences, Universitat Internacional de Catalunya, Barcelona, Spain; [2]Department of Methodology and ERI-Lectura, Universitat de València, Valencia, Spain; [3]Nebrija Research Center in Cognition, Universidad Antonio de Nebrija, Madrid, Spain and [4]Department of Information and Communication Technologies, Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain

Research Article

Cite this article: Cordero, G., Paredes-Paredes, J.R., Perea, M., Sebastian-Galles, N. and Díaz, B. (2025). Voice processing ability predicts second-language phoneme learning in early bilingual adults. *Bilingualism: Language and Cognition* **28**, 1209–1226. https://doi.org/10.1017/S136672892400110X

Received: 16 July 2024
Revised: 4 November 2024
Accepted: 3 December 2024
First published online: 17 January 2025

Keywords:
voice processing; phoneme learning; individual differences; confirmatory factor analysis; structural equation modeling

Corresponding author:
Gaël Cordero;
Email: gcordero@uic.es

This article has earned badges for transparent research practices: Open Data and Open Materials. For details see the Data Availability Statement.

Abstract

Individuals differ greatly in their ability to learn the sounds of second languages, even when learning starts early in life. Recent research has suggested that the ability to identify the idiosyncratic acoustic variations introduced into the speech stream by the speaker might be relevant for second-language (L2) phoneme learning. However, only a positive correlation between voice recognition and phoneme learning has been shown. In the present study, we investigated whether voice processing ability predicts L2 phoneme learning. We employed a battery of behavioral cognitive ability measures to assess voice processing ability and L2 phoneme learning in 57 early bilingual adults. Confirmatory factor analyses (CFAs) and structural equation modeling (SEM) revealed that voice processing ability predicts L2 phoneme learning. Our findings align with theories of speech perception that attribute a fundamental role to the analysis of voice cues and suggest that the accurate identification of speaker-specific variation is also relevant for phoneme learning.

Highlights

- High individual differences in voice processing and L2 phoneme learning
- CFAs support voice processing and L2 phoneme learning being distinct abilities
- SEMs of accuracy and reaction time data show that voice ability predicts L2 phoneme learning

1. Introduction

Anyone who has taken a second-language (L2) course will have noticed that we display considerable individual differences in language learning. Some people struggle with the most basic abilities, while others seem to absorb linguistic knowledge effortlessly. One of the most challenging aspects of learning an L2 is the acquisition of its speech sounds (i.e., phonemes), an ability subject to great individual differences, with only a minority of learners achieving high proficiency (Schmitz et al., 2018; Sebastian-Galles & Baus, 2005; Sebastian-Galles & Díaz, 2012). Studies have found that individual differences in L2 phoneme command persist despite accounting for comparable experiences and opportunities to learn the L2 (Archila-Suerte et al., 2016; Díaz et al., 2012; Sebastian-Galles & Baus, 2005; Sebastian-Galles & Díaz, 2012). Yet, the learner-related factors that impact L2 phoneme command are poorly understood. A recent study (Díaz et al., 2022) showed that individual differences in L2 phoneme proficiency were related to the ability to recognize trained (i.e., learned) voices. Here, we tested whether voice processing abilities (operationalized as the ability to recognize and discriminate voices) can predict attained L2 phoneme learning in a sample of early bilingual adults using a battery of behavioral tests and structural equation models (SEMs).

Speech is a highly variable and complex signal. It contains both linguistic information, which reflects the message the speaker intends to transmit, and voice information, which provides cues about various characteristics of the speaker. Listeners use linguistic information to understand what is being said, while voice information is exploited for successful social interactions (Nygaard & Tzeng, 2021). The complexity and variability of the speech signal are largely due to these two types of information not being discreetly encoded; there is no one-to-one mapping between the percepts of phonemes and their acoustic correlates across speakers (Peterson & Barney, 1952). The anatomy of the vocal tract, which is responsible for speech production, is unique to each speaker. Consequently, the acoustic characteristics of each speaker's voice are also unique. The

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

main acoustic features that characterize voices are the average fundamental frequency, which is perceived as voice pitch, and the frequency values of formants (i.e., resonances of the vocal tract), that cause the percept of vocal timbre (Baumann & Belin, 2010; Ghazanfar & Rendall, 2008; Latinus & Belin, 2011). While the first and second formants (F_1 and F_2) are claimed to be the primary cues to determine vowel identity (Fox et al., 1995; Yang & Fox, 2014), higher formants have been proposed to carry most of the vocal timbre information, as they exhibit minimal within-speaker variation across vocalizations (Kitamura & Akagi, 1995). However, as stated, the spectral values of all formants are determined by the anatomy of the speaker's vocal tract. Therefore, theories of speech perception must address how the perceptual system resolves the lack of invariance between speech-sound (i.e., phoneme) percepts and their acoustic correlates across speakers.

Many of the solutions proposed by speech perception theories to this lack of invariance problem require the accurate identification of the speaker-specific spectro-temporal changes embedded in the speech signal. Speaker normalization theories argue that speech perception is accomplished by initially identifying the acoustic idiosyncrasies introduced into the speech stream by the speaker and discarding them from further processing. Thus, only the phoneme cues that enable the recognition of the corresponding phoneme representations are retained (Choi et al., 2018; Johnson & Sjerps, 2021; Nusbaum & Magnuson, 1997; Zhang & Chen, 2016). However, the specific acoustic cues onto which normalization is applied vary across theoretical proposals, such as the ratio between the F_1 and F_2 of vowels or the absolute fundamental frequency, and remain a matter of debate (for a review, see Persson & Jaeger, 2023). Conversely, distributional (Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011) and exemplar-based (Goldinger, 1998; Klatt, 1979; Sumner et al., 2014) models of speech perception do not consider voice information as noise to be discarded, but rather as fundamental for speech perception. These models propose that the speech perceptual system resolves the lack of invariance between phoneme percepts and their acoustic correlates by representing voice-dependent variations of speech. While distributional models claim that listeners retain statistical distributions of the range of variability of phoneme cues across speakers, exemplar-based models propose that listeners store memory traces of actual speech segments that contain both linguistic and voice details. Thus, according to these two theoretical proposals, speaker variations of phoneme productions are accounted by either inferring the most probable outcome or by a similarity matching process, respectively. Both distributional and exemplar-based models of speech perception share the underlying assumption that exposure to speaker variability provides the speech perceptual system the capability to accurately perceive speech. Numerous studies have repeatedly shown that voice and linguistic information interact during speech processing, as contemplated by all of the enumerated theoretical proposals. The perception of synthesized ambiguous vowels is strongly dependent on the spectro-temporal characteristics of a speaker's voice in a preceding sentence (Darwin et al., 1989; Krumbiegel et al., 2022; Ladefoged & Broadbent, 1957; Miller et al., 1984; Nearey, 1989; Newman & Sawusch, 2009; Reinisch & Sjerps, 2013; Sjerps et al., 2013, 2019) regardless of language familiarity (Sjerps & Smiljanić, 2013). Familiarity with a speaker is beneficial for speech comprehension in acoustically challenging scenarios, such as noisy environments or multi-talker situations (Drozdova et al., 2019; Johnsrude et al., 2013; Magnuson et al., 2021; Nygaard et al., 1994; Nygaard & Pisoni, 1998; Souza et al., 2013; Yonan & Sommers, 2000).

A growing body of evidence suggests that voice processing ability, the capacity of a listener to identify the speaker-specific acoustic variations introduced into the speech stream, is not only relevant for speech and speaker recognition (Johnson & Sjerps, 2021; Nygaard & Tzeng, 2021) but might also influence phoneme learning. The acquisition of non-native phonetic contrasts is enhanced if learnt from multiple speakers as compared to a single speaker (Bradlow et al., 1997; Bradlow & Pisoni, 1999; Deng et al., 2018; Iverson et al., 2005; Lively et al., 1993, 1994; Logan et al., 1991; Wong, 2014; Ylinen et al., 2010; Zhang et al., 2021, but see Brekelmans et al., 2022). This benefit in L2 phoneme learning is assumed to arise from the exposure to greater acoustic–phonetic variability that multiple speakers entail. This variability would allow L2 learners to identify the acoustic properties that convey linguistic information across speakers and facilitate accurate speech perception when new speakers are encountered (Deng et al., 2018; Iverson et al., 2005; Ylinen et al., 2010). The relevance of voice processing ability for language learning processes was also reported by Houston and Jusczyk (2000), who found that familiarity with characteristics of a speaker contributes to speech segmentation during early language learning. Infants were familiarized with isolated words spoken by one speaker and then presented with passages enunciated by a different speaker that occasionally contained the familiarized words. Seven-and-a-half-month-old infants recognized the trained words when familiarized and tested with speakers of the same sex but were unable to generalize across sexes. Houston and Jusczyk (2000) proposed that the ability to accurately disentangle voice information from linguistic information develops in parallel with language acquisition.

Additional evidence advocating for the importance of voice processing ability for language learning is provided by research in dyslexia, a developmental disorder characterized by difficulties in reading and spelling despite normal intelligence, neurological integrity, and educational opportunities. Current conceptualizations attribute dyslexia to an underlying phonological deficit that impedes the optimal association between phonemes and their respective characters (Ramus, 2003). Behavioral studies have established an association between dyslexia and difficulties in voice recognition (Perea et al., 2014; Perrachione et al., 2011). Perea et al. (2014) found that children and adults with dyslexia exhibited an impairment to recognize speakers in both the language for which they had previous phoneme representations, i.e., their native language (L1), and an unfamiliar language, leading them to suggest that poor voice recognition skill is a trait of dyslexia (Perea et al., 2014, but see Perrachione et al., 2011). This interpretation is in line with electrophysiological work that showed that children with dyslexia exhibit a reduced encoding of features related to pitch as compared to typically developing children (Chandrasekaran et al., 2009) and suggests that deficient voice processing ability might underlie the phonological deficit that characterizes dyslexia.

Further evidence that suggests that phoneme learning and voice processing are related abilities is provided by the advantage in voice recognition bilinguals exhibit compared to monolinguals when discriminating speakers in an unfamiliar language (Fecher & Johnson, 2019, 2022; Levi, 2019). Fecher and Johnson (2019, 2022) proposed that a richer phonetic upbringing had given rise to bilingual infants possessing higher sensitivity to phonetic cues, thus facilitating speaker recognition despite the absence of reliable phoneme representations. While a richer phonetic upbringing may underlie bilinguals having an advantage in voice recognition over monolinguals, bilingualism cannot account for the positive correlation between individual differences in voice recognition and

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

L2 phoneme learning a recent study observed, since the sample was entirely composed of early bilingual adults with similar opportunities to learn the L2 (Díaz et al., 2022). This study took advantage of the considerable variance displayed by Spanish (L1)–Catalan (L2) early bilinguals in their capacity to discriminate the Catalan-specific vowel contrast /e/ - /ɛ/, since native Spanish speakers perceive both phonemes as the Spanish vowel /e/ (Bosch et al., 2000; Pallier et al., 1997, 2001; Sebastian-Galles et al., 2006; Sebastian-Galles & Soto-Faraco, 1999). This phenomenon, where two L2 speech sounds are perceived as a single phoneme from the native language, is known as *perceptual assimilation* and constitutes one of the most challenging scenarios L2 speakers face (Best & Tyler, 2007; Flege, 1995). The bilinguals studied by Díaz et al. (2022) were selected from a previous study (Schmitz et al., 2018) according to whether they had exhibited either native-like or below-native performance in three behavioral tasks that evaluated their ability to perceive the L2-specific vowel contrast /e/ - /ɛ/. The bilinguals were administered a voice recognition task (adapted from Perea et al, 2014; Perrachione et al., 2011), which required them to learn associations of voices speaking in the participants' first language and cartoon avatars while their behavioral and electroencephalographic responses were registered.

In addition to the voice recognition task, Díaz et al. (2022) administered a non-word association task (NWAT) which required participants to learn associations between auditory non-words enunciated by a single speaker and cartoon avatars. The task served to obtain a behavioral measure of the participants' general capacity to learn audiovisual associations, an ability that might have influenced participants' performance in the voice recognition task. The behavioral data showed that voice recognition ability positively correlated with attained L2 phoneme discrimination, while none of these two measures correlated with NWAT. Analysis of the electroencephalographic data revealed a positive correlation between the brain activity during voice recognition and the behavioral L2 phoneme discrimination ability at two time windows: 300–340 and 880–1140 ms. These findings were in line with previous studies, which had reported voice recognition eliciting positive brain electrophysiological responses 300 ms after stimuli onset (Humble et al., 2019; Schweinberger, 2001; Zäske et al., 2014, 2018). The positive relation between voice recognition (at the behavioral and electroencephalographic levels) and L2 phoneme discrimination ability evidenced a common individual variance for L2 phoneme and voice recognition processes. The new-found relation between these two seemingly independent processes opened up the possibility of voice processing abilities impacting the final attainment of L2 phonemes. Díaz et al. (2022) suggested that the correlation between voice recognition ability and L2 phoneme learning might stem from L2 learners with proficient voice processing skills being better equipped to disentangle voice and linguistic information during learning, resulting in finer-tuned L2 phoneme representations and thus greater accuracy when detecting L2 phonemes. However, this proposal was limited by the correlational nature of the evidence.

In the present study, we examined if the ability to accurately identify the acoustic idiosyncrasies introduced into the speech stream by a speaker (i.e., voice processing ability) predicts L2 phoneme learning using structural equation modeling (SEM, for a list of all acronyms used in this article, see Appendix 1). We employed a battery of behavioral tests to assess voice processing ability and attained L2 phoneme learning in a sample of 57 early Spanish (L1)–Catalan (L2) bilingual adults with similar characteristics as the participants in Díaz et al. (2022). Voice processing ability was operationalized as the ability to recognize and discriminate speakers. We assessed participants' voice recognition skills using three different tasks. The first of these three was a voice recognition task in the native language (L1) of the participants, Spanish, which was identical to the task employed in Díaz et al. (2022). This L1 voice recognition task consisted in training participants to recognize five voices and subsequently testing voice recognition accuracy. Recognizing voices in one's L1 is facilitated by the prior phonological and semantic knowledge of the spoken language (Yu et al., 2023) and results in greater accuracy as compared to recognizing voices in an unknown language (Lx) (Perea et al., 2014; Perrachione et al., 2011). To obtain a richer characterization of the voice processing ability of participants than in Díaz et al. (2022), we also administered an Lx voice recognition task similar to the one employed by Perea et al. (2014) in which the voices spoke Chinese. By using both L1 and Lx voice recognition tasks, we aimed to capture participants' ability to identify voice cues that are intertwined with linguistic information in two different situations; when prior linguistic knowledge facilitated the identification of voice cues (L1 voice recognition task) and when prior linguistic knowledge did not facilitate voice recognition (Lx voice recognition task). Lastly, to deepen our understanding of voice processing abilities, we assessed participants' ability to identify speaker-specific cues embedded in the speech signal in the absence of linguistic-dependent acoustic variations. For this purpose, we designed a novel voice discrimination task (VDT) which required participants to evaluate whether two emotional interjections (Belin et al., 2008) had been produced by the same or different unfamiliar speakers. We employed affect bursts as stimuli due to emotional tone being a within-person source of non-linguistic variation which drastically modulates the spectro-temporal characteristics of the speech signal (Lavan et al., 2019a). These three voice tasks therefore evaluated participants' voice processing abilities in three situations that varied in their engagement of speech processes: linguistic information present and familiar (i.e., L1 voice recognition task), linguistic information present but unfamiliar (i.e., Lx voice recognition task), and linguistic information not present (voice discrimination task). The participants' L2 phoneme learning ability was quantified using two tasks that evaluated L2 phoneme knowledge at the sub-lexical and lexical levels, respectively: a categorization task (CT) of synthetic vowels (Pallier et al., 1997) and an auditory lexical decision task (Schmitz et al., 2018; Sebastian-Galles et al., 2005; Sebastian-Galles & Baus, 2005). All tasks measured accuracy and reaction time (RT). While both accuracy scores and RT capture effective cognitive processing, they are qualitatively different measures. Accuracy scores capture how similar the decision alternatives are to each other and how effectively the correct option can be identified. RT measures the speed with which a participant identifies the correct option. Perceptual decision-making models have highlighted the need to study both measures when investigating individual differences since, surprisingly, they tend to exhibit low correlation on an individual level (Ratcliff et al., 2010; Ratcliff et al., 2015a, 2015b). Drawing firm conclusions in behavioral studies therefore necessitates interpreting both measures (Ratcliff et al., 2015a).

We conducted confirmatory factor analysis (CFA) to investigate whether both the accuracy and RT data, modeled separately, were represented more adequately by two related latent variables (i.e., voice processing ability and L2 phoneme learning), as hypothesized, or rather by a single latent variable (i.e., general speech ability). After confirming that the model with two latent variables provided an overall better fit of the data, we proceeded to

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

investigate our main hypothesis that voice processing ability predicted L2 phoneme learning with SEM. A positive result would provide insight into the high variability early bilingual adults display in their command of L2 phonemes and suggest that voice processing influences L2 learning.

2. Methods

2.1. Sample size estimation

The minimum sample size required for this study was estimated using an *a priori* power analysis (Hancock & Mueller, 2013). Using a tool designed for SEM studies (Soper, 2023), we calculated the minimum sample size as a function of the number of observed and latent variables (5 and 2, respectively), anticipated effect size (β = .61, based on Díaz et al., 2022), desired probability (p = .05) and statistical power (π = .80). This analysis determined that a minimum of 12 participants was necessary to detect an effect. However, to ensure the convergence of the CFAs and SEMs, we aimed to collect the data of a minimum of 50 participants, following the recommendation of Bentler and Chou (1987) of having a minimum of 10 participants per indicator.

2.2. Participants

The sample of this study was composed of 57 Spanish–Catalan bilingual adults (40 female; mean age 21 years; age range 18–26) born and raised in the metropolitan area of Barcelona in Catalunya, an autonomous community of Spain where Spanish and Catalan are co-official languages. The L1 of all participants was Spanish; they had been raised in monolingual Spanish families and had not been systematically exposed to Catalan until the age of 4 years, when mandatory bilingual schooling begins. All participants were highly fluent speakers of Catalan; from kindergarten on, they had received mandatory bilingual education. At the time of testing, all participants were pursuing or had obtained a university degree in Catalonia, indicating that they had completed mandatory bilingual schooling, a requirement to access higher education.

Participants were selected using an online survey in Google Forms that collected information concerning their personal history (place of birth, place/s of residence, etc.) and language profile (L1, L2, age of acquisition of each spoken language, current use of each spoken language, etc.) of the respondent and their extended family. This was done to ensure that the participants had no substantial experience with any language other than Spanish during their initial years of life (0–4 years of age) and that systematic exposure to Catalan only began upon commencing mandatory bilingual schooling. Participants answered free-response questions inquiring about the language(s) employed to communicate with each family in their early childhood environment. All participants reported exclusively communicating in Spanish with both of their parents and other regular caretakers. None of the participants had extended family members or caretakers from the eastern region of Andalusia nor the Region of Murcia, two autonomous communities in the south of Spain. This was avoided because the Spanish dialects in these regions employ the phoneme /ɛ/ and the standard Spanish /e/ (Sanders, 1994; Soriano, 2012). Participants exposed to one of these Spanish dialects during their early infancy would have had an advantage in distinguishing the two phonemes we exploited to evaluate L2 phoneme learning.

None of the participants possessed substantial musical training, as defined by a previous study (Kaganovich et al., 2013). Substantial musical training consisted in meeting a minimum of two of the three following criteria: (1) the onset of musical training having occurred before the age of 12 years; (2) having partaken in musical training for a minimum of 5 years; and (3) being part of a musical group or ensemble, either currently or in the past. None of the participants had received a clinical diagnosis of a hearing problem, learning disability, or neurological impairment. Of the 1123 respondents that completed the online questionnaire, only 68 were eligible for inclusion in the final sample, of which 57 accepted to participate in this study. Participants provided their written informed consent and were monetarily compensated for their time (10 €). The Medical Faculty and Health Sciences Ethics Committee of the Universitat Internacional de Catalunya approved the procedures (Protocol no.º: PSI-2020-05).

2.3. Materials

A battery composed of six behavioral tasks was employed to evaluate the participants' voice processing ability, L2 phoneme learning, and general audiovisual learning capacities. Voice processing ability was assessed with the L1 voice recognition task (L1 VRT), the Lx voice recognition task (Lx VRT), and the VDT. The indicators of L2 phoneme learning were a CT and a lexical decision task (LDT). General audiovisual learning was evaluated with the NWAT. All tasks registered both accuracy and RT data and were programmed and executed in MATLAB (Version R2021a, MathWorks, Inc., Natick, MA USA) using the Psychophysics Toolbox extensions (3.0.18; Brainard, 1997; Pelli, 1997). Here, we present a summarized description of the tasks. A detailed description can be found in the Supplementary Materials.

2.3.1. Voice recognition tasks (VRTs)

The L1 VRT and the Lx VRT, adapted from Perea et al. (2014), followed an identical procedure. These two tasks solely differed in the stimuli they employed. In the L1 VRT, the auditory stimuli consisted of 10 Spanish sentences recorded by 5 Spanish native speakers, while 10 Chinese sentences recorded by 5 Chinese native speakers were employed in the Lx VRT. Ten female avatars were created, of which five were employed in each VRT. The VRTs trained participants to associate voices with avatars and then tested the learning that the participants had attained. Participants were taught the associations between voices and avatars in two phases, each composed of 25 trials: the training and the short test. The trials of the training followed an ABX structure; two voice–avatar pairings were sequentially presented. One of the two voices was then repeated while the five avatars were displayed. Participants had to indicate as fast as possible by means of a button press which of the five avatars the repeated voice corresponded to. Feedback was provided concerning the participants' response accuracy, and the correct avatar was displayed on the screen. The trials of the short test consisted in the presentation of an auditory stimulus accompanied by the five avatars. Participants indicated as fast as they could which of the five avatars was associated with the presented voice. As in the training, feedback was provided after each delivered response. The test phase, composed of 50 trials, followed the same structure as the short test but no feedback was provided. Participants were trained and tested on different sentences.

2.3.2. Voice discrimination task (VDT)

The Montreal Affective Voices set (Belin et al., 2008) was employed as the stimuli of the VDT. This set is composed of 10 different speakers enunciating nine affective interjections using the vowel /ɑ/.

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

The VDT followed an AX discrimination design: Two auditory stimuli were sequentially presented and participants indicated via button press as fast as they could if the same or different speakers had enunciated the two vocalizations. In half of the trials, both stimuli had been enunciated by the same speaker, while in the other half, they had been enunciated by different speakers. Fifty-two trials composed the VDT.

2.3.3. Categorization task (CT)

The CT followed the design presented by Pallier and collaborators (1997). The stimuli consisted of a continuum of seven synthesized vowel stimuli between the Catalan vowels /e/ and /ɛ/. In 63 trials (nine trials per stimuli), participants had to respond as fast as they could via button press if the vowel they heard was perceived as the first vowel in the Catalan word Pere (/perə/, the name Peter) or as the first vowel in pera (/pɛrə/, which means pear).

2.3.4. Lexical decision task (LDT)

The LDT employed in this study was from Sebastian-Galles et al. (2005). The stimuli consisted of 344 auditory stimuli (experimental and control) enunciated by a native Catalan speaker. The experimental stimuli included 132 words containing one of the two phonemes from the targeted Catalan contrast (i.e., /e/ or /ɛ/) and 132 non-words which were designed by substituting the /e/ and /ɛ/ present in the real words with the other member of the phoneme pair. Eighty control stimuli, 40 Catalan words and 40 non-words were also employed. Control non-words were derived from a set of Catalan words different from the control and experimental words. These control non-words were created by changing a vowel phoneme in this separate set of Catalan words with a phoneme employed in both Spanish and Catalan. In each of the 212 trials, participants were presented with an auditory stimulus and had to respond via button press if the stimulus was part of the Catalan lexicon. The experimental stimuli were distributed between two lists to ensure that participants only heard one member of the same word pair. Both lists included all control stimuli, and their use was counterbalanced across participants.

2.3.5. Non-word association task (NWAT)

The NWAT was initially introduced in Díaz et al. (2022). Six non-words enunciated by a single native Spanish speaker constituted the auditory stimuli for this task while six avatars constituted the visual stimuli. The NWAT sought to train and test participants' ability to learn audiovisual associations. It was composed of two phases: a training and a test. Each of the 12 trials of the training phase consisted in the simultaneous presentation of a non-word–avatar pairing. The test trials, a total of 48, consisted of the presentation of a non-word, while the six avatars were displayed. Participants indicated via button press which avatar was associated with the presented non-word as fast as possible.

2.4. Procedure

The six tasks were administered in a single one-hour-long experimental session. The tasks were presented to all participants in the following order: Lx VRT, LDT, VDT, L1 VRT, NWAT, and, lastly, the CT. The order of task presentation was arbitrary; however, the order of the tasks was maintained constant throughout for participants to avoid task-order effects playing a role in individual task performance. Instructions for each task were displayed via text. Any doubts the participants had were resolved by the experimenters before commencing each task. Instructions were delivered in Catalan for the LDT and the CT and in Spanish for the other four tasks. Participants were instructed to provide their responses with their dominant hand and to keep their response fingers over the response buttons. For all participants, the six tasks were presented on an HP EliteBook 840 G7 Notebook PC with Audio-Technica ATH-PRO7x headphones, ensuring a consistent and comfortable audio level. Participants were tested individually in sound-attenuated rooms at the Psychology and Psychiatry University Clinic and Digital Media Studios of the Universitat Internacional de Catalunya and at the laboratories of the Center for Brain and Cognition of the Pompeu Fabra University.

2.5. Data analysis

We investigated whether voice processing ability predicted L2 phoneme learning using SEM, a statistical methodology that systematically analyzes the relationship among several variables. Following Brown (2015), CFAs were conducted prior to the SEMs. CFA assesses the relationships between observed measures and latent variables. CFA allows for the validation of the hypothesized latent constructs being manifested through the employed indicators. Similar to a previous study (Díaz et al., 2022), we tested whether general audiovisual learning abilities influenced the participants' performance in the VRTs by computing Pearson's correlations between the accuracy scores and RT of the NWAT and the VRTs. Mplus Version 8.8. Demo (Mplus. Statistical Analysis with Latent Variables, 2017) was used to estimate the CFAs and SEMs. All other analyses were conducted with R 4.2.2 (R Core Team, 2019) and RStudio 2022.12.0 (RStudio Team, 2020).

Each task's accuracy and RT scores were computed from trials where participants delivered their responses within a specific time window. These time windows were designed to exclude responses provided before perceptual processing while including responses delivered up to three-and-a-half seconds after mean stimuli duration, similar to one of our previous studies (Sebastian-Galles et al., 2005). The time windows for each task were as follows: L1 VRTs: 250–7500 ms; Lx VRT: 250–8500 ms; VDT: 250–5000 ms; CT: 250–4000 ms; LD: 250–4000 ms; and NWAT: 250–4000 ms. Following these criteria, the following percentage of data was discarded for each task: L1 VRT: 0.70%; Lx VRT: 3.12%; VDT: 0.67%; CT: 3.07%; LD: 2.53%; and NWAT: 5.52%. Due to technical malfunctions, the LDT data of two participants were not registered. Under the assumption of data missing at random, multiple imputations by chained equations were performed with the R package mice. Subsequently, multivariate normality was assessed using the Mahalanobis distance (D^2_M) and computed with the R stat function. D^2_M was calculated for each participant's responses to the five experimental tasks, and its statistical significance was tested with χ^2 at a significant level of .001 (Kline, 2015).

Accuracy scores were computed for each participant and each task. For the VRTs, we computed the proportion of accurate responses delivered, following studies which have previously employed voice recognition tasks (Díaz et al., 2022; Perea et al., 2014; Perrachione et al., 2011). For the VDT (see Table A1 in Appendix 2 for descriptive statistics of the proportion of correct responses), since it aimed to evaluate the ability of participants to discriminate between pairs of stimuli, we computed the d', an index of discriminability (see Table A2 in Appendix 3 for mean proportion of hits and false alarms) derived from signal detection theory (McNicol, 2005; Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). Accuracy scores for the CT were computed as in previous studies (Schmitz et al., 2018; Sebastian-Galles et al., 2005;

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

Sebastian-Galles & Baus, 2005). We sought to obtain a measure that reflected if participants could perceive the difference between the /e/ stimuli (steps 1 and 2) and the /ɛ/ stimuli (steps 6 and 7). For this, the average /e/ responses to steps 6 and 7 were subtracted from the average /e/ responses of steps 1 and 2. Thus, high positive scores reflect a good separation of /e/ and /ɛ/, scores close to zero reflect that participants did not respond differently to steps 1 and 2 than to steps 6 and 7, and negative scores indicate that participants' responses showed a reverse pattern. Negative CT scores were assumed to originate from responses systematically delivered in reverse, which necessitates the capacity to perceive the difference between phoneme categories. Thus, the CT scores were transformed into absolute values. For the LDT, the mean accuracy for the experimental words was computed (see Table A1 in Appendix 2). Previous studies that had used the same LDT with the same population had computed the A' score, a non-parametric unbiased index of sensitivity (McNicol, 2005; Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999), due to the participant's strong bias to consider most experimental non-words as real words (Schmitz et al., 2018; Sebastian-Galles et al., 2005; Sebastian-Galles & Baus, 2005). After confirming that our participants showed a high rate of false alarms for the experimental stimuli of the LDT (see Table A2 in Appendix 3), consistent with previous studies (Schmitz et al., 2018; Sebastian-Galles et al., 2005; Sebastian-Galles & Baus, 2005), we computed A' scores for the LDT. A' ranges between a score of 0.5 (random response) and 1.0 (perfect discrimination). To ensure high L2 lexical knowledge, we excluded participants with an A' < 0.8 in the control trials of the LDT. Lastly, for the NWAT, we employed the proportion of accurate responses as the accuracy score, following the study in which this task was introduced (Díaz et al., 2022). RT scores for all tasks and participants resulted from the mean average of the RT corresponding to trials in which the correct response was delivered.

Separate CFAs and SEMs were constructed for the accuracy scores and the RT data. The model parameters of the SEMs and CFAs were estimated using the robust maximum likelihood estimator, which does not rely on the assumption of a normal distribution (Kline, 2015). While theoretically the tasks we employed are indicators of two different, yet related, constructs (i.e., voice processing ability and L2 phoneme learning), we also tested the possibility of a single-latent variable model providing an adequate fit of the data to rule out a possible explicative model that might be supported statistically. We employed the Akaike information criterion (AIC) to compare between the models with two latent variables (i.e., voice processing ability and L2 phoneme learning) and the models with a single latent variable (i.e., general speech ability) (Akaike, 1998). The chi-square test of model fit (χ^2) was considered significant at $p < .05$. A significant result of this statistic would indicate model misfit, reflecting a deviation between the population covariance structure and the model-implied covariance structure (Kline, 2015). Goodness of fit of the models was also assessed via two indices which are robust in models with relatively small degrees of freedom, as in the present study (Shi et al., 2022): the comparative fit index (CFI) and the standardized root mean residual (SRMR). CFI compares the fit of the specified model to a baseline null model in which the latent variables are unrelated by constraining the covariance between the latent variables to zero. SRMR represents the average squared deviation between the observed and reproduced covariances. Following the recommendation of Hu and Bentler (1999), the following values in the indices were interpreted as indicating a good fit: CFI ≥ .90 and SRMR ≤ .08. For completeness, we report the root mean square error of approximation (RMSEA), a measure of model misfit due to model misspecification commonly employed in models with large degrees of freedom, though not recommended for models with small degrees of freedom as those presented here (Kenny et al., 2015).

3. Results

3.1. Tasks' results

All indicators exhibited considerable variability, suggesting that the tasks we employed successfully captured individual differences (Figures 1 and 2). The skewness and kurtosis values were within the thresholds suggested by Hancock and Mueller (2013) (i.e., absolute values of 2 and 7, respectively) for conducting CFAs and SEMs (see Table 1). Covariance matrices were generated (see Tables 2 and 3) as part of the standard procedure of conducting CFAs and SEMs (Kline, 2015). All participants attained high accuracy scores in the control trials of the LDT (M = 0.95; SD = 0.04; range = 0.83–0.99), and therefore, no participant was excluded from the analysis due to having low L2 lexical knowledge. Multivariate normality was assessed using D^2_M to rule out the possibility of disturbances caused by potential multivariate outliers. No multivariate outliers were identified for the accuracy scores, and all participants were included in the CFAs and SEMs. For the RT data, a single case was identified as a multivariate outlier following the D^2_M criteria and was excluded from the RT models.

We did not expect performance in the NWAT (accuracy mean = 0.7; accuracy SD = 0.26; RT mean = 1589.15 ms; RT SD = 252.89 ms) to correlate significantly with performance in the VRTs, since these tasks were designed to capture individual differences of different abilities (i.e., general audiovisual learning and voice recognition abilities, respectively). No correlation between these tasks was observed in a previous study (Díaz et al., 2022). We ascertained that individual differences in general audiovisual learning abilities were not related to performance in the VRTs by computing Pearson's correlation coefficients between the scores of the VRTs and those of the NWAT. Performance in the NWAT did not correlate with L1 VRT measures (accuracy: $r = .16$; $p = .221$; RT: $r = .19$; $p = .151$) nor Lx VRT (accuracy: $r = .14$; $p = .298$; RT: $r = .13$; $p = .336$), suggesting that individual differences in general audiovisual learning abilities were not related to performance in the VRTs. As a result, the NWAT data was not included in subsequent analyses. Given the dominance of female participants in our sample, we ascertained that gender did not influence participants' performance in the indicators of voice processing ability and L2 phoneme learning using a series of Welch's t-tests for unequal sample sizes. No comparison between genders approached statistical significance (all ps > 0.1; see Table A3 in Appendix 4).

3.2. Confirmatory factor analyses (CFAs)

CFAs were computed to evaluate whether the accuracy scores and RT data captured the latent constructs as intended. We tested whether voice processing ability and L2 phoneme learning could be modeled as distinct but related constructs. Additionally, we modeled the data into a single-latent-variable structure to test the possibility of this competing model. The CFAs with two related latent variables (see Figure 3) showed that the accuracy scores in the L1 VRT and Lx VRT were valid indicators of voice processing ability (both $p < .001$). While VDT accuracy did not significantly represent voice processing ability ($p = .191$), RT in this same task

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

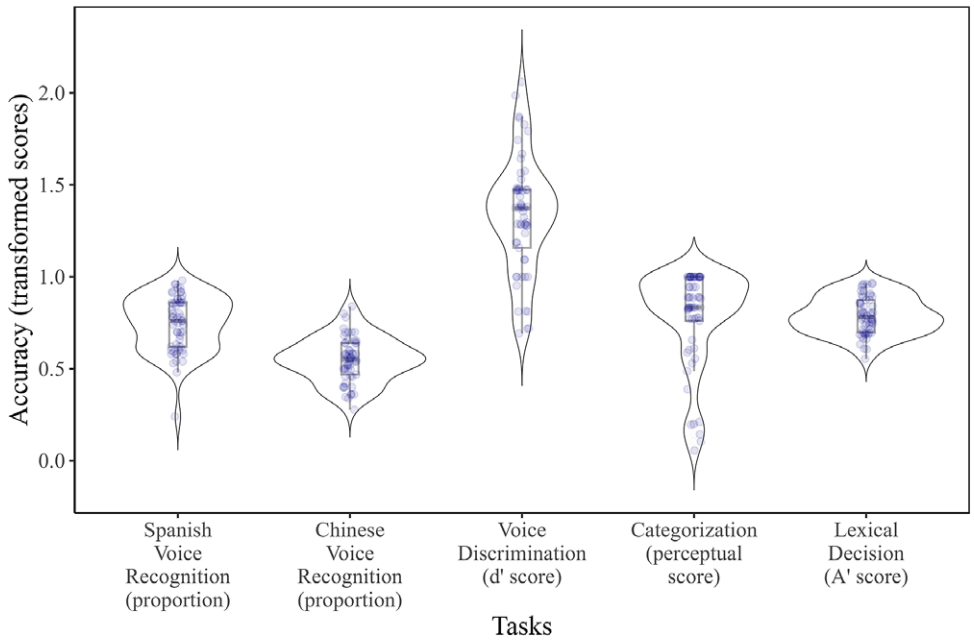


Figure 1. Accuracy scores of the indicators of voice processing ability (Spanish voice recognition, Chinese voice recognition, and voice discrimination) and L2 phoneme learning (categorization and lexical decision). Note that different accuracy transformed scores are depicted and direct visual comparison between the tasks is discouraged.

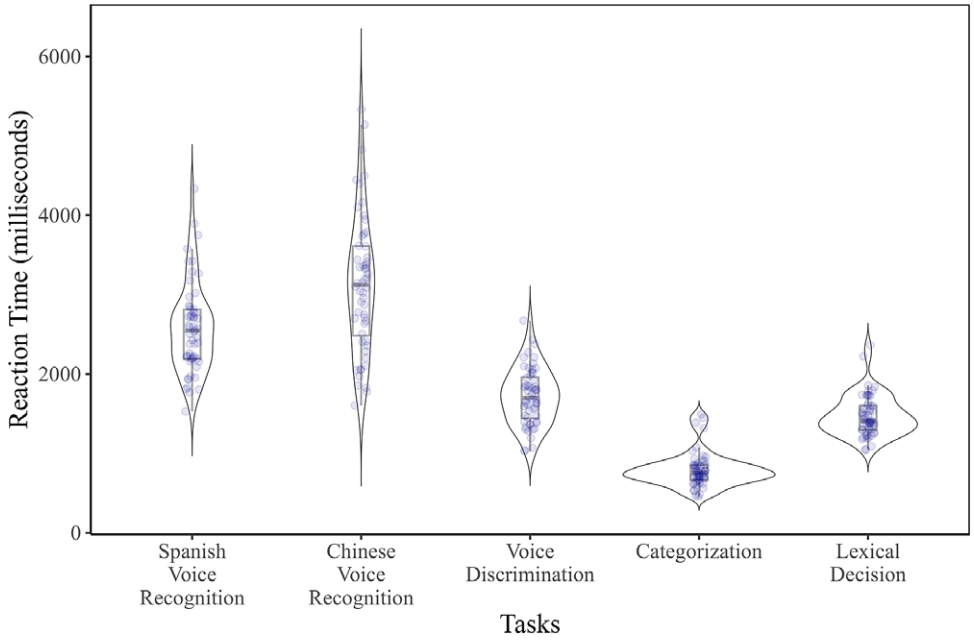


Figure 2. RT for the indicators for voice processing ability (Spanish voice recognition, Chinese voice recognition, and voice discrimination) and L2 phoneme learning (categorization and lexical decision).

did ($p < .001$). Furthermore, RT in L1 VRT and Lx VRT represented voice processing ability ($p < .001$). Concerning L2 phoneme learning, both the accuracy scores and RT in the CT and the LDT represented this latent construct (all $p < .001$). Voice processing ability and L2 phoneme learning were correlated in both the accuracy and the RT model. The chi-square test of model fit (χ^2) was not significant for either CFA, indicating that the models with two related latent variables provided an adequate fit of the data. The CFI and SRMR indicated that both the accuracy CFA and the RT CFA met the established criteria for goodness of fit (Hu and Bentler, 1999) (see Table 4).

The value of the χ^2 test indicated that the single-latent CFA modeled with the accuracy scores (see Figure 4) fitted the data adequately ($p > .05$). However, the single-latent CFA modeled with the RT data exhibited significant model misfit ($\chi^2(5) = 13.368$; $p < .05$), suggesting that the model could not adequately represent the data (see Table 4). All accuracy scores of all tasks significantly represented general speech ability ($p < .005$), with the sole exception

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

Table 1. Descriptive statistics of the accuracy scores and reaction times of the indicators

Task	M	SD	Min.	Max.	Skewness	Kurtosis
Accuracy (n = 57)						
Spanish voice recognition (hits)	0.75	0.15	0.24	0.98	−0.66	0.38
Chinese voice recognition (hits)	0.55	0.12	0.28	0.84	0.03	−0.37
Voice discrimination (d')	1.33	0.32	0.69	2.06	−0.02	−0.28
Categorization (perceptual score)	0.78	0.26	0.06	1.00	−1.38	0.87
Lexical decision (A')	0.79	0.10	0.55	0.97	−0.05	−0.87
Reaction time (n = 56)						
Spanish voice recognition	2574	568	1526	4333	0.78	0.55
Chinese voice recognition	3126	861	1603	5332	0.43	−0.25
Voice discrimination	1717	361	1027	2672	0.24	−0.34
Categorization	791	219	454	1492	1.42	2.32
Lexical decision	1474	275	1044	2364	1.18	1.57

Indicators of voice processing ability are Spanish voice recognition, Chinese voice recognition, and voice discrimination. Indicators of L2 phoneme learning are categorization and lexical decision. ms = milliseconds.

Table 2. Covariance matrix of the accuracy score data

Task	1	2	3	4	5
1. Spanish voice recognition (hits)	–				
2. Chinese voice recognition (hits)	.0033	–			
3. Voice discrimination (d')	.0133	.0028	–		
4. Categorization (perceptual score)	.0191	.0112	.0130	–	
5. Lexical decision (A')	.0056	.0013	.0022	.0128	–

Table 3. Covariance matrix of the reaction time data (all indicators presented in ms).

Task	1	2	3	4	5
1. Spanish voice recognition	–				
2. Chinese voice recognition	361320.48	–			
3. Voice discrimination	109375.28	17983080			
4. Categorization	26987.40	63014.83	32787.74		
5. Lexical decision	55461.89	114627.04	50248.20	2441.09	–

of the VDT (p = .139). Fit indicators for this single-latent CFA exhibited adequate fit results following the criterion suggested by Hu and Bentler (1999). Comparison of the Akaike Information Criterion (AIC) for models based on accuracy scores suggested that the CFA with a single latent variable provided a more adequate representation of the accuracy scores than the CFA with two latent variables (see Table 4). However, the single-latent CFA model did not adequately fit the RT data while the CFAs with two latent variables showed adequate fit for both the accuracy scores and the RT data. Hence, modeling voice processing ability and L2 phoneme learning as distinct but related constructs provided an overall better characterization of the complete dataset.

3.3. Structural equation models (SEMs)

We investigated whether voice processing ability predicted L2 phoneme learning with SEMs. The similarity between the procedures of the VRTs motivated us to release the covariate parameter between them when estimating the models. The results of the SEM analyses were in line with CFA findings. For both the accuracy and RT models, all measures of voice processing ability loaded onto said factor, with the sole exception of the VDT accuracy, with a loading that was close to significance (p = .066). Both accuracy and RT measures of L2 phoneme learning are loaded onto an L2 phoneme learning factor. Voice processing ability predicted L2 phoneme learning in both the model that included the accuracy scores (p < .005) and the model that included the RT data (p < .001) (see Figure 5). The goodness-of-fit indicators employed to evaluate the models met the criteria proposed by Hu and Bentler (1999), indicating that the data were well represented by the models (see Table 5).

4. Discussion

We investigated whether individual differences in voice processing ability provided a statistically significant prediction regarding L2 phoneme learning proficiency. To test this hypothesis, we exploited the variance Spanish (L1)–Catalan (L2) early bilinguals display in their capacity to discriminate the Catalan-specific vowel contrast /e/ - /ɛ/. We employed a battery of behavioral tests to assess voice processing ability and L2 phoneme learning in a sample of 57 early bilingual adults. Performance in all indicators exhibited considerable variability, suggesting that the tasks we employed successfully captured individual differences. We employed CFA to evaluate whether the accuracy scores and RT data captured two distinct latent constructs, as hypothesized, or a single latent variable. The model with two related latent variables showed a good fit of both the accuracy and RT data while the model with a single latent variable only fitted the accuracy data. Subsequent SEMs incorporating two latent variables for both accuracy scores and RT data confirmed that voice processing ability is a reliable predictor of L2 phoneme learning in early bilingual adults. Drawing on various theories of

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

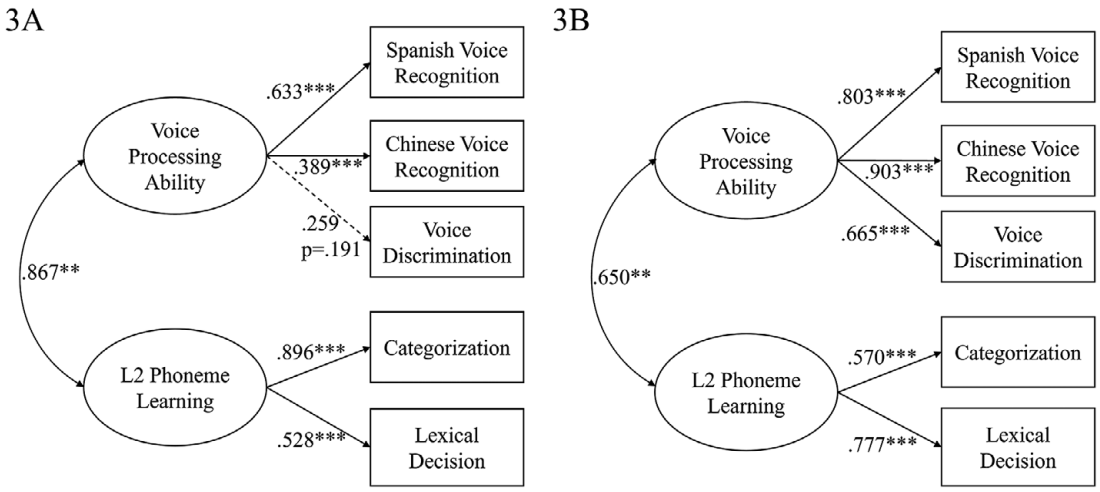


Figure 3. Accuracy (3A) and RT (3B) CFAs with two correlated latent variables. Paths connecting the latent variables (circles) are the correlations between these constructs. The values between the latent variables and the manifest variables (squares) represent the standardized loadings of each task onto the latent variable. All loadings were significant at the $p < .001$ except for voice discrimination (VDT) to voice processing ability ($p = .191$) in the accuracy CFA. $^{**}p < .01$, $^{***}p < .001$.

Table 4. Goodness-of-fit indices' results of the CFAs

CFA model	χ^2	*df*	*P*	CFI	SRMR	AIC	RMSEA	RMSEA CI 90%
Accuracy model/two latent variables	4.271	4	.371	.992	.042	−197.599	.034	(.000; .206)
Accuracy model/single latent variable	3.719	5	.591	1	.046	−199.300	0	(.000; .158)
Response time model/two latent variables	7.097	4	.865	.996	.051	4085.090	.118	(.000; .256)
Response time model/single latent variable	13.368	5	.020	.909	.071	4088.689	.173	(.063; .288)

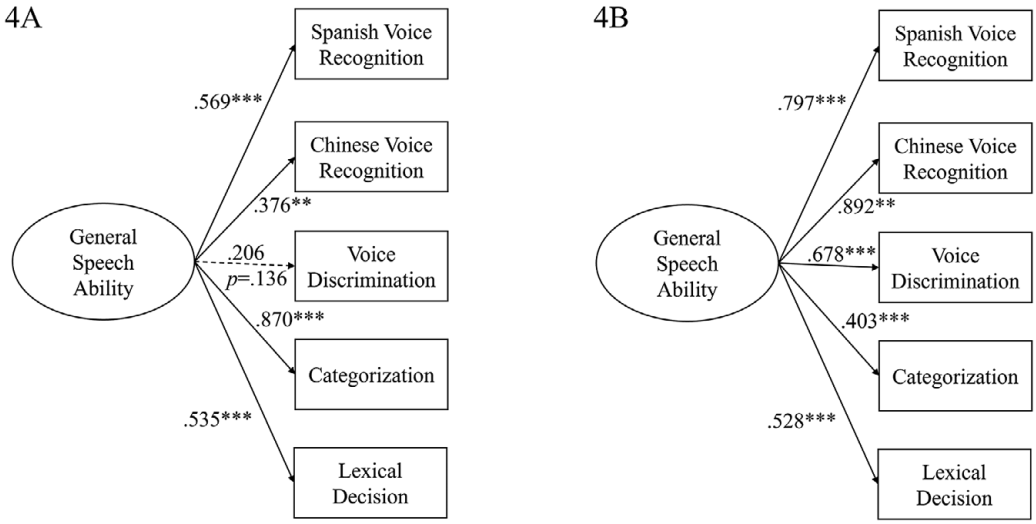


Figure 4. Accuracy (4A) and RT (4B) CFAs with a single latent variable. We present 4B for informational purposes only, since the RT data are misrepresented by this model (see Table 4). The values between the latent variable (circle) and the manifest variables (squares) represent the standardized loadings of each task onto the latent variable. All loadings were significant except for voice discrimination (VDT) in the accuracy CFA ($p = .136$). ** $p < .01$, *** $p < .001$.

speech perception, in the following paragraphs, we discuss the nature of the relationship between voice processing and L2 phoneme learning. We also consider how voice processing abilities may relate to language learning in different stages of life, such as learning an L2 as an adult and acquiring a native language. Furthermore, we offer some considerations for future studies that seek to further investigate the influence of voice processing abilities on language learning.

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

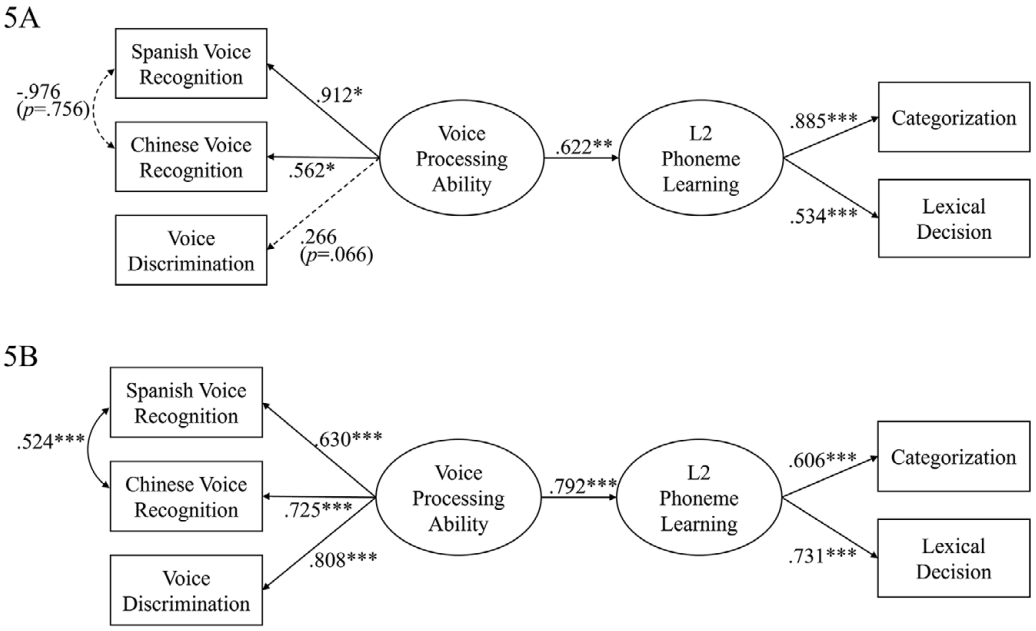


Figure 5. Accuracy (5A) and RT (5B) SEMs showing the effect from the latent variable voice processing ability over L2 phoneme learning. The values between the latent variables (circles) and their respective manifest variables (squares) represent the standardized loadings of each task onto the corresponding latent variable. All loadings were significant, except for VDT to voice processing ability in 5A, which approached significance (p = .066). A dashed line represents non-significant results. * p < .05, ***p < .001.

Table 5. Goodness-of-fit indices' results of the accuracy and RT SEMs

Model	χ^2	*df*	*p*	CFI	SRMR	RMSEA
Accuracy model	1.810	3	.613	1	.030	0
Reaction time model	2.368	3	.500	1	.023	0

Our findings suggest that the ability of a listener to identify the idiosyncratic acoustic variations introduced into the speech stream by the speaker's voice, an ability that theoretical proposals of native speech perception consider indispensable (Johnson & Sjerps, 2021; Nygaard & Tzeng, 2021), relates to L2 phoneme learning ability. It should be noted that theoretical models of non-native speech perception do not address how non-native listeners cope with the lack of invariance of speech sounds across speakers (Best, 1994; Best & Tyler, 2007; Escudero, 2009; Flege, 1995). Models of non-native speech perception assume that to create representations of non-native phonemes, the speech perception system needs to identify the invariant phonemic cues that differentiate these from native phonemes. These models implicitly assume that the mechanisms that enable L1 perception are the same that support the identification of the phonemic cues that distinguish native from non-native phonemes. We therefore build on native speech perception models to examine the potential mechanisms that drive the present relation between voice and L2 phoneme abilities.

Being models of native speech perception, speaker normalization theories do not address non-native phoneme learning. However, we provide a tentative explanation of how this theoretical proposal might accommodate the association between voice processes and L2 phoneme learning abilities. Speaker normalization theories propose that to account for the high variability of the speech signal, the perceptual system initially identifies and discards the voice information embedded in the speech signal. This computation entails that the remaining acoustic information cannot be attributed to speaker idiosyncrasies but rather corresponds to linguistic information (Choi et al., 2018; Johnson & Sjerps, 2021; Nusbaum & Magnuson, 1997; Zhang & Chen, 2016). Viewed through the theoretical frame of speaker normalization theories, individual differences in voice processing abilities might be relevant during L2 phoneme learning as they would determine the listener's accuracy in identifying the spectro-temporal correlates of voices in the speech signal, such as variations of the fundamental frequency or the frequency of the formants (Baumann & Belin, 2010; Ghazanfar & Rendall, 2008; Latinus & Belin, 2011). Inadequate identification of speaker-specific acoustic variation could lead to two potential scenarios: the speech system would either flag phoneme-relevant cues as voice-dependent and discard them from speech analyses or rather consider voice cues as phoneme-relevant features and include them in further speech processing. In both cases, inaccurate identification of voice features would hamper the discovery of the invariant cues of non-native phonemes and their subsequent learning. A caveat to this interpretation lies in the nature of the computations assessed in our voice processing ability tasks. The voice tasks focus on explicit recognition and discrimination and might involve high-level processes, such as accessing identity representations or making similarity judgments. These high-level processes may differ from those that underpin speaker normalization, which is typically conceptualized as an automatic process that mostly relies on low-level acoustic contrasts (Sjerps et al., 2013; Sjerps & Smiljanić, 2013).

An alternative theoretical proposal which also accommodates interactions between voice and linguistic information is provided by distributional and exemplar-based models of speech perception. These models suggest that, rather than a normalization process occurring, the speech perceptual system tracks and retains speaker-specific acoustic variations introduced into the speech signal (Goldinger, 1998; Klatt, 1979; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011; Sumner et al., 2014). The flexibility that these models attribute to speech perception, conceptualizing it

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

as a dynamic ability capable of incorporating novel information to adapt to new scenarios (e.g., learning dialectal variations), can arguably accommodate the learning of non-native phoneme contrasts. Based on phonetic training paradigms that show greater generalization when learning occurs in multispeaker as compared to single speaker conditions, it has been proposed that the speech system dynamically learns and extrapolates the features that characterize phonemes across speakers (Weatherholtz & Jaeger, 2016). Building on distributional and exemplar-based models, the association between voice processing ability and L2 phoneme learning might originate from the listener's ability to properly identify the speaker-specific variations introduced in the speech signal, directly impacting the ability of the listener to discover the acoustic correlates of phonemic regularities. The tasks employed in the present study to measure voice processing ability are designed to capture both low-level and high-level acoustic processes, similar to the processes conceptualized by distributional and exemplar-based models. It should be noted that these models propose that learning regular variations of voices is an implicit process, while the behavioral tasks employed in this study evaluated explicit learning and discrimination. However, recent research has shown that voice recognition accuracy is similar regardless of whether attention is directed to the voice or to the linguistic content of speech, suggesting that both implicit and explicit processes support the learning of the relevant cues that characterize voices (Lee & Perrachione, 2022). While the assumptions of these models fit well with the reported findings, the validity of these theories of speech perception remains a subject of ongoing debate. Therefore, we are limited with regard to drawing a causal interpretation from the observed predictive value of voice processing ability for L2 phoneme learning. Investigating the neural underpinnings that support the interaction between voice processing ability and L2 phoneme learning may further our understanding of the relation between these two processes, especially upon considering that speaker recognition and speech perception engage partially distinct brain regions (Bonte et al., 2014; Formisano et al., 2008; Schall et al., 2015). Previous studies have proposed two neurofunctional mechanisms that might support interactions between voice and speech processes: (i) interhemispheric functional connectivity between right lateralized voice-sensitive regions and left lateralized speech-sensitive regions (Deng et al., 2018; Kreitewolf et al., 2014; von Kriegstein et al., 2010) and (ii) the functional overlap exhibited by regions along the temporal cortices and right temporoparietal junction, which exhibit sensitivity to both voice and phonetic information (Chandrasekaran et al., 2011; Formisano et al., 2008; Holmes & Johnsrude, 2021; Luthra et al., 2023; Myers & Theodore, 2017; von Kriegstein et al., 2010). If these mechanisms are also engaged during L2 phoneme learning, they would provide a neural basis for the interaction between voice processing ability and L2 phoneme learning that would align with the proposals of the models of speech perception that we have discussed here.

Despite the present findings fitting well with theoretical proposals, it remains unknown whether the predictive value of voice processing abilities for L2 phoneme learning can be extrapolated to learning during other stages of life. The participants in this study were early bilingual adults who learnt the L2 upon commencing mandatory bilingual schooling at the age of 4 years. While children predominantly utilize implicit domain-specific mechanisms in language learning, adult L2 learners can no longer rely on these implicit mechanisms. Instead, they must reflect on the grammatical structure of the novel language and exploit general cognitive strategies (DeKeyser, 2000). Furthermore, recent studies support the long-standing proposal of the existence of a sensitive period for language learning (Hartshorne et al., 2018; Werker & Hensch, 2015). Sensitive periods are developmental stages during which the central nervous system exhibits greater experience-induced plasticity, enabling the acquisition of sensory and cognitive abilities. Once a sensitive period has ended, poorer learning is possible in that domain. Crucially, the bilinguals tested in the present study learnt the L2 after the sensitive period for phoneme learning had concluded, which has been proposed to end during the second year of life (for a review, see Werker & Hensch, 2015). Indeed, several studies show that systematic exposure to an L2 at the age at which our sample of participants began learning does not consistently result in native-like proficiency in L2 phoneme contrast discrimination, as would be expected if the L2 had been acquired during the sensitive period (Caramazza et al., 1973; Díaz et al., 2012; Schmitz et al., 2018; Sebastian-Galles & Díaz, 2012). Therefore, the observed association between voice processing and L2 phoneme learning may generalize to the learning of non-native phoneme contrasts occurring after the sensitive period for phoneme acquisition concludes. Supporting this claim, previous research has shown that voice processes are relevant for language learning during adulthood. For instance, numerous studies have demonstrated significant gains in the perception of L2 phoneme contrasts when learners are exposed to these contrasts from multiple speakers, as compared to learning from a single speaker (Bradlow et al., 1997; Bradlow & Pisoni, 1999; Deng et al., 2018; Iverson et al., 2005; Lively et al., 1993, 1994; Logan et al., 1991; Wong, 2014; Ylinen et al., 2010; for a review, see Zhang et al., 2021). This benefit in L2 phoneme learning in multispeaker contexts is believed to reflect the enhanced identification of the invariant cues that characterize phonemes when the learner has access to a more diverse speech input (Deng et al., 2018; Iverson et al., 2005; Ylinen et al., 2010). However, it remains to be investigated whether adult learners display variability in their ability to extract the features that characterize phonemes across speakers and whether this variability is related to individual differences in voice processing ability.

The assessment of voice abilities may be relevant to predict not only phoneme learning in the L2 but also the acquisition of the L1. Previous studies (Perea et al., 2014; Perrachione et al., 2011) revealed an association between difficulties in voice recognition and dyslexia, a difficulty in learning to read whose origins are claimed to be rooted in a phonological deficit (Ramus, 2003). Impaired voice recognition abilities have been proposed as a marker of developmental dyslexia and a valuable measure to predict the disability (Perea et al., 2014). Moreover, an electrophysiological study reported a reduced encoding of features related to pitch in children with dyslexia compared to typically developing children (Chandrasekaran et al., 2009). Chandrasekaran et al. (2009) suggested that individuals with dyslexia may experience challenges adapting speech processes to accommodate the characteristics of different voices. Considering voice processing as a general mechanism that enables the learning of speech sound invariants would provide an explanatory mechanism for the co-occurrence in dyslexia of voice and phoneme deficits. However, extrapolating an effect that influences L2 phoneme learning to the acquisition of the L1 would require further testing. The neural processes that enable language learning during the first years of life are different than those that enable learning after that sensitive period has concluded (Hartshorne et al., 2018; Werker & Hensch, 2015). Furthermore, theoretical models of non-native speech perception conceptualize the acquisition of the L2 as qualitatively different from the learning of an L1, since L2 learners must identify the cues

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

that differentiate non-native from the native phonemes (Best, 1994; Best & Tyler, 2007; Escudero, 2009; Flege, 1995). Thus, investigating whether voice processing ability influences L1 phoneme learning would also shed light on the similarities and differences between learning an L1 and an L2.

The discussed implications of the present findings for language learning call for further research to better comprehend the nature of the relationship between voice processing abilities and L2 phoneme learning. Future studies that investigate how voice processing ability influences language learning should note that our battery of behavioral tests captured large individual differences in L2 phoneme proficiency in both sub-lexical and lexical contexts, as reported in previous studies that investigated similar populations with the same L2 phoneme tasks (Díaz et al., 2012; Schmitz et al., 2018; Sebastian-Galles et al., 2005; Sebastian-Galles & Díaz, 2012). We were also successful in replicating the high inter-individual variability in the ability to recognize and discriminate speakers that previous studies observed in healthy populations (Aglieri et al., 2017; Lavan et al., 2019a; Mühl et al., 2018). While previous studies evaluated voice processing with speech samples containing phonetic information from the participants' native language, we employed a diverse set of experimental procedures to evaluate voice abilities in the participants' native language, in an unfamiliar language, and from sub-lexical affect bursts. We observed variability in all indicators of voice processing ability, regardless of the participants' familiarity with the language employed during the voice tasks, whether the task trained participants to recognize the speaker or the linguistic content (sub-lexical or lexical) of the task. This suggests that, while they likely influence task performance, neither language familiarity, voice familiarity, nor linguistic content are critical factors when evaluating voice processing ability in healthy populations. However, we acknowledge that the accuracy data of the VDT did not relate to voice ability in the CFA. While all voice processing ability indicators captured individual differences, the VDT differed considerably from the other two voice tasks: It did not involve processing of linguistic information or training and employed affective interjections that primarily modulate the fundamental frequency of the speech signal (Bachorowski et al., 2001; Bachorowski & Owren, 2001; Lavan et al., 2016, 2019b), unlike phoneme changes, which primarily encoded as changes in the F1 and F2 (Fox et al., 1995; Yang & Fox, 2014). These three differences could explain why the VDT task did not relate to voice ability in the CFA for the accuracy data. If future research supports the idea that linguistic content is not a crucial factor to capture individual differences in voice processing ability, it could lead to the development of a voice-processing evaluative tool applicable to any population, regardless of their linguistic background.

The combined use of CFAs and SEM revealed that the proficiency early L2 learners achieve in mastering L2 phoneme contrasts, an ability known to vary considerably among individuals (Archila-Suerte et al., 2016; Díaz et al., 2012; Schmitz et al., 2018; Sebastian-Galles & Baus, 2005; Sebastian-Galles & Díaz, 2012), can be predicted based on an individual's ability to recognize and discriminate voices. Our models showed this effect despite the tasks employed as indicators of voice processing ability involving learning and memory components not present in the indicators of L2 phoneme learning. In other words, as noted by a reviewer, had the tasks employed as indicators of each construct been more similar in their domain-general cognitive requirements, the predictive capacity of voice processing ability on L2 phoneme learning would likely have been greater than that reported here. Furthermore, voice and phoneme processing differ in the relative importance of various acoustic features of the speech signal. Research suggests that voice processing is primarily dependent on changes at high spectral modulations (i.e., >1.1 cycles per octave at center frequencies of up to 0.8 kHz), while phoneme category is mostly determined by changes in lower spectral modulations (i.e., broad spectral modulations for center frequencies above 0.6 kHz) and fast temporal changes (i.e., >7.8 Hz) (Rutten et al., 2019). Therefore, the predictive capacity of voice processing ability over L2 phoneme learning is not due to both processes relying on the same acoustic features. However, this study does not establish a definitive causal relation between voice processing ability and L2 phoneme learning. While theoretical accounts of speech perception could support a causal relation, it remains feasible that the association between voice processing ability and L2 phoneme learning stems from a common origin: The listener's sensitivity to detect phoneme changes in any given language. This interpretation was also presented in the study that inspired the current investigation (Díaz et al., 2022) and is based on two sets of findings: speaker recognition accuracy being influenced by the phoneme knowledge of the listener (Fecher & Johnson, 2019, 2022; Perrachione et al., 2011) and the relation between the mastery of L2 phoneme contrasts with the ability to discriminate both native and unfamiliar phoneme contrasts (Díaz et al., 2008, 2016). However, the alternative interpretation of voice processing ability and L2 phoneme learning emerging from a common underlying process lacked conclusive support from the single-latent CFAs. The analysis yielded good fit for the accuracy data but failed to adequately fit of the RT data. Advocating for the validity of the single-latent model would entail disregarding the RT data, a measure of effective cognitive processing equally valid, and complementary, to accuracy data (Ratcliff et al., 2015a). Another potential limitation of the current study is the relatively small sample size. Some recommendations suggest employing sample sizes of up to several thousand individuals when conducting SEM (Kline, 2015; Schumacker & Lomax, 2010). A sample size of such proportions was unfeasible due to the strict inclusion criteria participants had to meet. Nonetheless, *a priori* power analysis confirmed that our analyses were sufficiently powered, and, indeed, both the CFA and SEM exhibited good fit when including two latent variables. A second potential limitation related to the sample of this study is the greater number of women participants compared to men. However, no significant performance differences between males and females in the indicators of either latent variable were observed. This finding suggests that the higher proportion of female participants did not influence our primary findings.

In conclusion, our findings contribute to understanding the processes involved in speech perception and language learning: Individual differences in voice processing ability among early bilingual adults can predict the proficiency they achieve in L2 phoneme learning. By recognizing voice processing as a predictive factor in language learning, we deepen our understanding of the variability in L2 proficiency observed among early bilingual adults. This perspective opens new avenues for research, ranging from the acquisition of the native language to educational applications.

Supplementary material. To view supplementary material for this article, please visit http://doi.org/10.1017/S136672892400110X.

Data availability statement. The data collected and the analysis code are accessible at https://osf.io/symg2/?view_only=9f331bf8fc2146099874e81de4e908ae.

Acknowledgements. This work was supported by grants from the Ministry of Science and Innovation of the Spanish Government, State Research Agency and

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

European Regional Development Fund (PID2019-106924GA-I00, PID2022-137368NB-I00 and PID2021-123416NB-I00 funded by MICIN/AEI/10.13039/501100011033/FEDER UE) awarded to BD and NSG. MP was awarded a grant from the Valencian Government (CIAICO/2021/172). NSG was awarded the ICREA Academia Prize by the Catalan Government. GC was supported by a doctoral fellowship from the Universitat Internacional de Catalunya. Two grants financed by the Catalan Generalitat AGAUR (2021 SGR 00911 and 2021 SGR 00625) also supported this work.

Competing interests. None declared.

References

Aglieri, V., **Watson, R.**, **Pernet, C.**, **Latinus, M.**, **Garrido, L.**, & **Belin, P.** (2017). The Glasgow voice memory test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, **49**, 97–110. https://doi.org/10.3758/s13428-015-0689-6

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer New York. https://doi.org/10.1007/978-1-4612-1694-0_15

Archila-Suerte, P., **Bunta, F.**, & **Hernandez, A. E.** (2016). Speech sound learning depends on individuals' ability, not just experience. *International Journal of Bilingualism*, **20**, 231–253. https://doi.org/10.1177/1367006914552206

Bachorowski, J.-A., & **Owren, M. J.** (2001). Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, **12**, 252–257. https://doi.org/10.1111/1467-9280.00346

Bachorowski, J.-A., **Smoski, M. J.**, & **Owren, M. J.** (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, **110**, 1581–1597. https://doi.org/10.1121/1.1391244

Baumann, O., & **Belin, P.** (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research Psychologische Forschung*, **74**, 110–120. https://doi.org/10.1007/s00426-008-0185-z

Belin, P., **Fillion-Bilodeau, S.**, & **Gosselin, F.** (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, **40**, 531–539. https://doi.org/10.3758/BRM.40.2.531

Bentler, P. M., & **Chou, C.-P.** (1987). Practical issues in structural modeling. *Sociological Methods & Research*, **16**, 78–117. https://doi.org/10.1177/0049124187016001004

Best, C. T. (1994). The emergence of native-language phonological in fluences in infants: A perceptual assimilation model. In J. C. Good man, & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). Cambridge: MIT Press.

Best, C. T., & **Tyler, M. D.** (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language learning & language teaching* (Vol. **17**, pp. 13–34). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.07bes

Bonte, M., **Hausfeld, L.**, **Scharke, W.**, **Valente, G.**, & **Formisano, E.** (2014). Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. *The Journal of Neuroscience*, **34**, 4548–4557. https://doi.org/10.1523/JNEUROSCI.4339-13.2014

Bosch, L., **Costa, A.**, & **Sebastian-Galles, N.** (2000). First and second language vowel perception in early bilinguals. *European Journal of Cognitive Psychology*, **12**, 189–221. https://doi.org/10.1080/09541446.2000.10590222

Bradlow, A. R., & **Pisoni, D. B.** (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, **106**, 2074–2085. https://doi.org/10.1121/1.427952

Bradlow, A. R., **Pisoni, D. B.**, **Akahane-Yamada, R.**, & **Tohkura, Y.** (1997). Training Japanese listeners to identify English /**r**/ and /**l**/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, **101**, 2299–2310. https://doi.org/10.1121/1.418276

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis*, **10**, 433–6. PMID: 9176952.

Brekelmans, G., **Lavan, N.**, **Saito, H.**, **Clayards, M.**, & **Wonnacott, E.** (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, **126**, 104352. https://doi.org/10.1016/j.jml.2022.104352

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. (2nd ed.). Guilford Publications.

Caramazza, A., **Yeni-Komshian, G. H.**, **Zurif, E. B.**, & **Carbone, E.** (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *The Journal of the Acoustical Society of America*, **54**, 421–428. https://doi.org/10.1121/1.1913594

Chandrasekaran, B., **Chan, A. H. D.**, & **Wong, P. C. M.** (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, **23**, 2690–2700. https://doi.org/10.1162/jocn.2011.21631

Chandrasekaran, B., **Hornickel, J.**, **Skoe, E.**, **Nicol, T.**, & **Kraus, N.** (2009). Context-dependent encoding in the human auditory brainstem relates to hearing speech in noise: Implications for developmental dyslexia. *Neuron*, **64**, 311–319. https://doi.org/10.1016/j.neuron.2009.10.006

Choi, J. Y., **Hu, E. R.**, & **Perrachione, T. K.** (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics*, **80**, 784–797. https://doi.org/10.3758/s13414-017-1395-5

Darwin, C. J., **Denis McKeown, J.**, & **Kirby, D.** (1989). Perceptual compensation for transmission channel and speaker effects on vowel quality. *Speech Communication*, **8**, 221–234. https://doi.org/10.1016/0167-6393(89)90003-4

DeKeyser, R. M. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, **22**, 499–533. https://doi.org/10.1017/S0272263100004022

Deng, Z., **Chandrasekaran, B.**, **Wang, S.**, & **Wong, P. C. M.** (2018). Training-induced brain activation and functional connectivity differentiate multi-talker and single-talker speech training. *Neurobiology of Learning and Memory*, **151**, 1–9. https://doi.org/10.1016/j.nlm.2018.03.009

Díaz, B., **Baus, C.**, **Escera, C.**, **Costa, A.**, & **Sebastian-Galles, N.** (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proceedings of the National Academy of Sciences*, **105**, 16083–16088. https://doi.org/10.1073/pnas.0805022105

Díaz, B., **Cordero, G.**, **Hoogendoorn, J.**, & **Sebastian-Galles, N.** (2022). Second-language phoneme learning positively relates to voice recognition abilities in the native language: Evidence from behavior and brain potentials. *Frontiers in Psychology*, **13**, 1008963. https://doi.org/10.3389/fpsyg.2022.1008963

Díaz, B., **Mitterer, H.**, **Broersma, M.**, **Escera, C.**, & **Sebastian-Galles, N.** (2016). Variability in L2 phonemic learning originates from speech-specific capabilities: An MMN study on late bilinguals. *Bilingualism: Language and Cognition*, **19**, 955–970. https://doi.org/10.1017/S1366728915000450

Díaz, B., **Mitterer, H.**, **Broersma, M.**, & **Sebastian-Galles, N.** (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, **22**, 680–689. https://doi.org/10.1016/j.lindif.2012.05.005

Drozdova, P., **van Hout, R.**, & **Scharenborg, O.** (2019). Talker-familiarity benefit in non-native recognition memory and word identification: The role of listening conditions and proficiency. *Attention, Perception, & Psychophysics*, **81**, 1675–1697. https://doi.org/10.3758/s13414-018-01657-5

Escudero, P. (2009). The linguistic perception of SIMILAR L2 sounds. In P. Boersma & S. Hamann (Eds.), *Phonology in perception* (pp. 151–190). De Gruyter Mouton. https://doi.org/10.1515/9783110219234.151

Fecher, N., & **Johnson, E. K.** (2019). Bilingual infants excel at foreign-language talker recognition. *Developmental Science*, **22**, e12778. https://doi.org/10.1111/desc.12778

Fecher, N., & **Johnson, E. K.** (2022). Revisiting the talker recognition advantage in bilingual infants. *Journal of Experimental Child Psychology*, **214**, 105276. https://doi.org/10.1016/j.jecp.2021.105276

Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Baltimore: York Press.

Formisano, E., **De Martino, F.**, **Bonte, M.**, & **Goebel, R.** (2008). «Who» is saying «what»? Brain-based decoding of human voice and speech. *Science*, **322**, 970–973. https://doi.org/10.1126/science.1164318

Fox, R. A., **Flege, J. E.**, & **Munro, M. J.** (1995). The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional

scaling analysis. *The Journal of the Acoustical Society of America*, **97**, 2540–2551. https://doi.org/10.1121/1.411974

Ghazanfar, A. A., & **Rendall, D.** (2008). Evolution of human vocal production. *Current Biology*, **18**, R457–R460. https://doi.org/10.1016/j.cub.2008.03.030

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, **105**, 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Hancock, G. R., & **Mueller, R. O.** (Eds.). (2013). *Structural equation modeling: A second course* (2nd ed). Information Age Publishing, Inc.

Hartshorne, J. K., **Tenenbaum, J. B.**, & **Pinker, S.** (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, **177**, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Holmes, E., & **Johnsrude, I. S.** (2021). Speech-evoked brain activity is more robust to competing speech when it is spoken by someone familiar. *NeuroImage*, **237**, 118107. https://doi.org/10.1016/j.neuroimage.2021.118107

Houston, D. M., & **Jusczyk, P. W.** (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, **26**, 1570–1582. https://doi.org/10.1037/0096-1523.26.5.1570

Hu, L., & **Bentler, P. M.** (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, **6**, 1–55. https://doi.org/10.1080/10705519909540118

Humble, D., **Schweinberger, S. R.**, **Dobel, C.**, & **Zäske, R.** (2019). Voices to remember: Comparing neural signatures of intentional and non-intentional voice learning and recognition. *Brain Research*, **1711**, 214–225. https://doi.org/10.1016/j.brainres.2019.01.028

Iverson, P., **Hazan, V.**, & **Bannister, K.** (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/−/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, **118**, 3267–3278. https://doi.org/10.1121/1.2062307

Johnson, K., & **Sjerps, M. J.** (2021). Speaker normalization in speech perception. In J.S. Pardo, L.C. Nygaard, R.E. Remez and D.B. Pisoni (Eds.), *The Handbook of Speech Perception* (pp. 145–176). Wiley. https://doi.org/10.1002/9781119184096.ch6

Johnsrude, I. S., **Mackey, A.**, **Hakyemez, H.**, **Alexander, E.**, **Trang, H. P.**, & **Carlyon, R. P.** (2013). Swinging at a cocktail party. *Psychological Science*, **24**, 1995–2004. https://doi.org/10.1177/0956797613482467

Kaganovich, N., **Kim, J.**, **Herring, C.**, **Schumaker, J.**, **MacPherson, M.**, & **Weber-Fox, C.** (2013). Musicians show general enhancement of complex sound encoding and better inhibition of irrelevant auditory change in music: An ERP study. *European Journal of Neuroscience*, **37**, 1295–1307. https://doi.org/10.1111/ejn.12110

Kenny, D. A., **Kaniskan, B.**, & **McCoach, D. B.** (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, **44**(3), 486–507. https://doi.org/10.1177/0049124114543236

Kitamura, T., & **Akagi, M.** (1995). Speaker individualities in speech spectral envelopes. *Journal of the Acoustical Society of Japan (E)*, **16**, 283–289. https://doi.org/10.1250/ast.16.283

Klatt, D. H. (1979). Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of Phonetics*, **7**, 279–312. https://doi.org/10.1016/S0095-4470(19)31059-9

Kleinschmidt, D. F., & **Jaeger, T. F.** (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, **122**, 148–203. https://doi.org/10.1037/a0038695

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. (4th ed.). Guilford publications.

Kreitewolf, J., **Friederici, A. D.**, & **von Kriegstein, K.** (2014). Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *NeuroImage*, **102**, 332–344. https://doi.org/10.1016/j.neuroimage.2014.07.038

Krumbiegel, J., **Ufer, C.**, & **Blank, H.** (2022). Influence of voice properties on vowel perception depends on speaker context. *The Journal of the Acoustical Society of America*, **152**, 820–834. https://doi.org/10.1121/10.0013363

Ladefoged, P., & **Broadbent, D. E.** (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, **29**, 98–104. https://doi.org/10.1121/1.1908694

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

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

Lavan, N., **Burton, A. M.**, **Scott, S. K.**, & **McGettigan, C.** (2019b). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, **26**, 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lavan, N., **Scott, S. K.**, & **McGettigan, C.** (2016). Laugh like you mean it: Authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, **40**, 133–149. https://doi.org/10.1007/s10919-015-0222-8

Lee, J. J., & **Perrachione, T. K.** (2022). Implicit and explicit learning in talker identification. *Attention, Perception, & Psychophysics*, **84**, 2002–2015. https://doi.org/10.3758/s13414-022-02500-8

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, **10**, 1–15. https://doi.org/10.1002/wcs.1483

Lively, S. E., **Logan, J. S.**, & **Pisoni, D. B.** (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, **94**, 1242–1255. https://doi.org/10.1121/1.408177

Lively, S. E., **Pisoni, D. B.**, **Yamada, R. A.**, **Tohkura, Y.**, & **Yamada, T.** (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, **96**, 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., **Lively, S. E.**, & **Pisoni, D. B.** (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, **89**, 874–886. https://doi.org/10.1121/1.410149

Luthra, S., **Magnuson, J. S.**, & **Myers, E. B.** (2023). Right posterior temporal cortex supports integration of phonetic and talker information. *Neurobiology of Language*, **4**, 1–33. https://doi.org/10.1162/nol_a_00091

Magnuson, J. S., **Nusbaum, H. C.**, **Akahane-Yamada, R.**, & **Saltzman, D.** (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, **83**, 1842–1860. https://doi.org/10.3758/s13414-020-02203-y

McMurray, B., & **Jongman, A.** (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, **118**, 219–246. https://doi.org/10.1037/a0022325

McNicol, D. (2005). *A primer of signal detection theory*. Psychology Press.

Miller, J. L., **Aibel, I. L.**, & **Green, K.** (1984). On the nature of rate-dependent processing during phonetic perception. *Perception & Psychophysics*, **35**, 5–15. https://doi.org/10.3758/BF03205919

Mühl, C., **Sheil, O.**, **Jarutytė, L.**, & **Bestelmeyer, P. E. G.** (2018). The Bangor voice matching test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods*, **50**, 2184–2192. https://doi.org/10.3758/s13428-017-0985-4

Myers, E. B., & **Theodore, R. M.** (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, **165**, 33–44. https://doi.org/10.1016/j.bandl.2016.11.001

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, **85**, 2088–2113. https://doi.org/10.1121/1.397861

Newman, R. S., & **Sawusch, J. R.** (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, **37**, 46–65. https://doi.org/10.1016/j.wocn.2008.09.001

Nusbaum, H. C., & **Magnuson, J. S.** (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability and speech processing* (pp. 109–132). Academic Press/Elsevier.

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

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

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

Nygaard, L. C., & **Tzeng, C. Y.** (2021). Perceptual integration of linguistic and non-linguistic properties of speech. In J.S. Pardo, L.C. Nygaard, R.E. Remez and D.B. Pisoni (Eds.), *The Handbook of Speech Perception* (pp. 398–427). Wiley. https://doi.org/10.1002/9781119184096.ch15

Pallier, C., **Bosch, L.**, & **Sebastian-Galles, N.** (1997). A limit on behavioral plasticity in speech perception. *Cognition*, **64**, B9–B17. https://doi.org/10.1016/S0010-0277(97)00030-9

Pallier, C., **Colomé, A.**, & **Sebastian-Galles, N.** (2001). The influence of native-language phonology on lexical access: Exemplar-based versus abstract lexical entries. *Psychological Science*, **12**, 445–449. https://doi.org/10.1111/1467-9280.00383

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, **10**, 437–442. https://doi.org/10.1163/156856897X00366

Perea, M., **Jiménez, M.**, **Suárez-Coalla, P.**, **Fernández, N.**, **Viña, C.**, & **Cuetos, F.** (2014). Ability for voice recognition is a marker for dyslexia in children. *Experimental Psychology*, **61**, 480–487. https://doi.org/10.1027/1618-3169/a000265

Perrachione, T. K., **Del Tufo, S. N.**, & **Gabrieli, J. D. E.** (2011). Human voice recognition depends on language ability. *Science*, **333**, 595–595. https://doi.org/10.1126/science.1207327

Persson, A., & **Jaeger, T. F.** (2023). Evaluating normalization accounts against the dense vowel space of Central Swedish. *Frontiers in Psychology*, **14**, 1165742. https://doi.org/10.3389/fpsyg.2023.1165742

Peterson, G. E., & **Barney, H. L.** (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, **24**, 175–184. https://doi.org/10.1121/1.1906875

R Core Team. (2019). *R: A language and environment for statistical computing* [Software]. https://www.r-project.org

Ramus, F. (2003). Developmental dyslexia: Specific phonological deficit or general sensorimotor dysfunction? *Current Opinion in Neurobiology*, **13**, 212–218. https://doi.org/10.1016/S0959-4388(03)00035-7

Ratcliff, R., **Smith, P. L.**, & **McKoon, G.** (2015a). Modeling regularities in response time and accuracy data with the diffusion model. *Current Directions in Psychological Science*, **24**, 458–470. https://doi.org/10.1177/0963721415596228

Ratcliff, R., **Thapar, A.**, & **McKoon, G.** (2010). Individual differences, aging, and IQ in two-choice tasks. *Cognitive Psychology*, **60**, 127–157. https://doi.org/10.1016/j.cogpsych.2009.09.001

Ratcliff, R., **Thompson, C. A.**, & **McKoon, G.** (2015b). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, **137**, 115–136. https://doi.org/10.1016/j.cognition.2014.12.004

Reinisch, E., & **Sjerps, M. J.** (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, **41**, 101–116. https://doi.org/10.1016/j.wocn.2013.01.002

RStudio Team. (2020). *RStudio: Integrated development for R.* RStudio [Software]. http://www.rstudio.com/.

Rutten, S., **Santoro, R.**, **Hervais-Adelman, A.**, **Formisano, E.**, & **Golestani, N.** (2019). Cortical encoding of speech enhances task-relevant acoustic information. *Nature Human Behaviour*, **3**, 974–987. https://doi.org/10.1038/s41562-019-0648-9

Sanders, B. P. (1994). *Andalusian vocalism and related processes* [Doctoral dissertation, University of Illinois at Urbana-Champaign]. https://doi.org/10.1016/j.jaci.2012.05.050

Schall, S., **Kiebel, S. J.**, **Maess, B.**, & **von Kriegstein, K.** (2015). Voice identity recognition: Functional division of the right STS and its behavioral relevance. *Journal of Cognitive Neuroscience*, **27**, 280–291. https://doi.org/10.1162/jocn_a_00707

Schmitz, J., **Díaz, B.**, **Fernández Rubio, K.**, & **Sebastian-Galles, N.** (2018). Exploring the relationship between speech perception and production across phonological processes, language familiarity, and sensory modalities. *Language, Cognition and Neuroscience*, **33**, 527–546. https://doi.org/10.1080/23273798.2017.1390142

Schumacker, R. E., & **Lomax, R. G.** (2010). *A beginner's guide to structural equation modeling* (3rd ed). Routledge.

Schweinberger, S. R. (2001). Human brain potential correlates of voice priming and voice recognition. *Neuropsychologia*, **39**, 921–936. https://doi.org/10.1016/S0028-3932(01)00023-9

Sebastian-Galles, N., & **Baus, C.** (2005). On the relationship between perception and production in L2 categories. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 279–292). Erlbaum.

Sebastian-Galles, N., & **Díaz, B.** (2012). First and second language speech perception: Graded learning. *Language Learning*, **62**, 131–147. https://doi.org/10.1111/j.1467-9922.2012.00709.x

Sebastian-Galles, N., **Echeverría, S.**, & **Bosch, L.** (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *Journal of Memory and Language*, **52**, 240–255. https://doi.org/10.1016/j.jml.2004.11.001

Sebastian-Galles, N., **Rodríguez-Fornells, A.**, **de Diego-Balaguer, R.**, & **Díaz, B.** (2006). First- and second-language phonological representations in the mental lexicon. *Journal of Cognitive Neuroscience*, **18**, 1277–1291. https://doi.org/10.1162/jocn.2006.18.8.1277

Sebastian-Galles, N., & **Soto-Faraco, S.** (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, **72**, 111–123. https://doi.org/10.1016/S0010-0277(99)00024-4

Shi, D., **DiStefano, C.**, **Maydeu-Olivares, A.**, & **Lee, T.** (2022). Evaluating SEM model fit with small degrees of freedom. *Multivariate Behavioral Research*, **57**, 179–207. https://doi.org/10.1080/00273171.2020.1868965

Sjerps, M. J., **Fox, N. P.**, **Johnson, K.**, & **Chang, E. F.** (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications*, **10**, 2465. https://doi.org/10.1038/s41467-019-10365-z

Sjerps, M. J., **McQueen, J. M.**, & **Mitterer, H.** (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, & Psychophysics*, **75**, 576–587. https://doi.org/10.3758/s13414-012-0408-7

Sjerps, M. J., & **Smiljanić, R.** (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics*, **41**, 145–155. https://doi.org/10.1016/j.wocn.2013.01.005

Snodgrass, J. G., & **Corwin, J.** (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, **117**(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Soper, D. S. (2023). *A-priori sample size calculator for structural equation models* [Software]. https://www.danielsoper.com/statcalc

Soriano, B. (2012). Andalusian vowel harmony and morphology-phonology interface. *Anuario del Seminario de Filología Vasca «Julio de Urquijo»*, **46**, 295–307. https://doi.org/10.1387/asju.12625

Souza, P., **Gehani, N.**, **Wright, R.**, & **McCloy, D.** (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, **24**, 689–700. https://doi.org/10.3766/jaaa.24.8.6

Stanislaw, H., & **Todorov, N.** (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, **31**, 137–149. https://doi.org/10.3758/BF03207704

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

von Kriegstein, K., **Smith, D. R. R.**, **Patterson, R. D.**, **Kiebel, S. J.**, & **Griffiths, T. D.** (2010). How the human brain recognizes speech in the context of changing speakers. *The Journal of Neuroscience*, **30**, 629–638. https://doi.org/10.1523/JNEUROSCI.2742-09.2010

Weatherholtz, K., & **Jaeger, T. F.** (2016). Speech perception and generalization across talkers and accents. In K. Weatherholtz & T. F. Jaeger, *Oxford research encyclopedia of linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.95

Werker, J. F., & **Hensch, T. K.** (2015). Critical periods in speech perception: New directions. *Annual Review of Psychology*, **66**, 173–196. https://doi.org/10.1146/annurev-psych-010814-015104

Wong, J. W. S. (2014). The effects of high and low variability phonetic training on the perception and production of English vowels /e/-/æ/ by Cantonese ESL

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

learners with high and low L2 proficiency levels. Interspeech 2014, 524–528. https://doi.org/10.21437/Interspeech.2014-129

Yang, J., & **Fox, R. A.** (2014). Perception of English vowels by bilingual Chinese–English and corresponding monolingual listeners. *Language and Speech*, **57**, 215–237. https://doi.org/10.1177/0023830913502774

Ylinen, S., **Uther, M.**, **Latvala, A.**, **Vepsäläinen, S.**, **Iverson, P.**, **Akahane-Yamada, R.**, & **Näätänen, R.** (2010). Training the brain to weight speech cues differently: A study of Finnish second-language users of English. *Journal of Cognitive Neuroscience*, **22**, 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Yonan, C. A., & **Sommers, M. S.** (2000). The effects of talker familiarity on spoken word identification in younger and older listeners. *Psychology and Aging*, **15**, 88–99. https://doi.org/10.1037/0882-7974.15.1.88

Yu, K., **Zhou, Y.**, **Zhang, L.**, **Li, L.**, **Li, P.**, & **Wang, R.** (2023). How different types of linguistic information impact voice perception: Evidence from the language-familiarity effect. *Language and Speech*, **66**, 1007–1029. https://doi.org/10.1177/00238309221143062

Zäske, R., **Limbach, K.**, **Schneider, D.**, **Skuk, V. G.**, **Dobel, C.**, **Guntinas-Lichius, O.**, & **Schweinberger, S. R.** (2018). Electrophysiological correlates of voice memory for young and old speakers in young and old listeners. *Neuropsychologia*, **116**, 215–227. https://doi.org/10.1016/j.neuropsychologia.2017.08.011

Zäske, R., **Volberg, G.**, **Kovacs, G.**, & **Schweinberger, S. R.** (2014). Electrophysiological correlates of voice learning and recognition. *Journal of Neuroscience*, **34**, 10821–10831. https://doi.org/10.1523/JNEUROSCI.0581-14.2014

Zhang, C., & **Chen, S.** (2016). Toward an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance*, **42**, 1252–1268. https://doi.org/10.1037/xhp0000216

Zhang, X., **Cheng, B.**, & **Zhang, Y.** (2021). The role of talker variability in nonnative phonetic learning: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, **64**, 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

Appendices

Appendix 1

All abbreviations employed in the present study in order of appearance.

L2	Second language
F_1	First formant
F_2	Second formant
L1	Native language
SEM	Structural equation model
Lx	Unfamiliar language
RT	Reaction time
CFA	Confirmatory factor analysis
VRT	Voice recognition task
VDT	Voice discrimination task
CT	Categorization task
LDT	Lexical decision task
NWAT	Non-word association task
AIC	Akaike's information criterion
χ^2	Chi-square test of model fit
CFI	Comparative fit index
SRMR	Standardized root mean residual
RMSEA	Root mean square error of approximation
D^2_M	Mahalanobis distance

Appendix 2

Table A1. Descriptive statistics of the proportion of accurate responses delivered to all experimental tasks

Task	M	SD	Min.	Max.	Skewness	Kurtosis
Spanish voice recognition	0.75	0.15	0.24	0.98	−0.66	0.38
Chinese voice recognition	0.55	0.12	0.28	0.84	0.03	−0.37
Voice discrimination	0.57	0.06	0.44	0.69	−0.11	−0.40
Categorization	0.57	0.41	0.00	1.00	−0.32	−1.70
Lexical decision	0.66	0.13	0.52	0.93	0.75	−0.84
Non-word association	0.70	0.26	0.20	1	−0.20	−1.43

n = 57 except for the lexical decision task in which n = 55.

Appendix 3

Table A2. Proportion of hits and false alarms for the VDT and LDT

Task	Hits (SD)	False alarms (SD)
VDT	0.58 (0.10)	0.45 (0.13)
LDT	0.95 (0.06)	0.63 (0.27)
LDT (control)	0.96 (0.04)	0.14 (0.12)

For the LDT, hits and false alarms have been calculated separately for experimental and control trials. Standard deviations are presented in brackets.

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press

Appendix 4

Table A3. Descriptive statistics and between-group comparisons as a function of sex for the indicators for voice processing ability and L2 phoneme learning

Task	Group		*t*	*df*	*p*-value
	Female	Male			
Accuracy (n = 57)					
Spanish voice recognition (hits)	0.75 (±0.13)	0.72 (±0.17)	0.75	24.37	.458
Chinese voice recognition (hits)	0.55 (±0.11)	0.54 (±0.14)	0.18	24.73	.860
Voice discrimination (d')	1.36 (±0.33)	1.28 (±0.26)	0.93	38.94	.359
Categorization (discrimination score)	0.78 (±0.24)	0.77 (±0.30)	0.15	25.69	.886
Lexical decision (A')	0.78 (±0.10)	0.80 (±0.10)	0.45	28.84	.655
Reaction time (n = 56)					
Spanish voice recognition (ms)	2626 (±601)	2456 (±483)	1.12	37.68	.271
Chinese voice recognition (ms)	3244 (±855)	2857 (±841)	1.57	31.01	.126
Voice discrimination (ms)	1716 (±327)	1722 (±441)	0.05	23.99	.96
Categorization (ms)	783 (±161)	810 (±322)	0.32	19.59	.75
Lexical decision (ms)	1473 (±224)	1471 (±338)	0.03	22.33	.976

ms = milliseconds.

https://doi.org/10.1017/S136672892400110X Published online by Cambridge University Press