Supplementary Materials of Voice Processing Ability Predicts Second-Language Phoneme Learning in Early Bilingual Adults.

Authors: Gaël Cordero, Jazmin R. Paredes-Paredes, Manuel Perea, Nuria Sebastian-Galles, Begoña Díaz

Detailed description of the tasks employed in the study.

Voice Recognition Tasks (VRTs)

The L1 VRT and the Lx VRT were identical, except for the stimuli each employed. These two tasks will be explained in a single section to reduce redundancy.

A total of ten female avatars were created using a free website ([www.crearunavatar.com](http://www.crearunavatar.com)), five were used in the L1 VRT, and the other 5 in the Lx VRT. The auditory stimuli employed in a previous study (Perea et al., 2014) were used. These stimuli consisted of twenty sentences: ten were in Spanish and employed in the L1 VRT, and the other ten were in Mandarin-Chinese and were employed in the Lx VRT. The Spanish sentences were recorded by five female native Spanish speakers (mean duration = 3000 ms; SD = 750 ms), and the Mandarin – Chinese sentences were recorded by five female native Mandarin speakers (mean duration = 4000 ms; SD = 800 ms). As in a previous study (Díaz et al., 2022), the VRTs comprised three phases: training, a short test, and a test.

The training phase aimed to teach participants to associate each of the five avatars with a speaker's voice. At the beginning of a trial, a fixation point was presented for 1 second. Subsequently, two different avatars were presented with an interstimulus interval (ISI) of 1s. During the ISI, a fixation point was displayed. Concurrent with the presentation of each avatar, the voice associated with that avatar enunciated a sentence. Only one sentence was employed per trial; thus, both voices enunciated the same sentence. Finally, the five avatars were presented with a number below (1- 5), and one of the two previous auditory stimuli was repeated. Participants were tasked with identifying the avatar associated with the voice by pressing the corresponding number on the PC keyboard as fast as possible. Feedback related to response accuracy was provided, as was the image of the correct avatar, which was displayed for 2 seconds at the end of each trial. The position of stimuli repetitions was evenly split, with half the trials repeating the stimuli associated with the first avatar and the other half the second. Only half of the sentences (5 in the L1 VRT and 5 in the Lx VRT) were employed in the training. The training phase included 25 trials (5 sentences x 5 avatars) and was followed by the short test phase. The short test phase sought to enhance the learning of the avatar–voice associations. Trials of the short test phase commenced with the presentation of a fixation point during 500ms. An auditory sentence was presented while the five avatars were displayed alongside their associated numbers. Participants were prompted to indicate via button press as quickly as possible which avatar was associated with the voice they had just heard, enunciating the sentence without the pressure of a time limit. After they had delivered their answer, feedback was provided concerning their response accuracy, and the correct avatar was displayed on the screen. Twenty-five trials (5 sentences x 5 avatars) composed the short test. Upon finalizing the short test, the test phase commenced. The test phase had the same design as the short test, with the critical differences that the sentences employed were novel to the participants and that no feedback was provided. Fifty trials (5 sentences x 5 avatars x 2 repetitions) comprised the test phase. Avatar-voice pairings were controlled via eight stimuli lists which were counterbalanced across participants. These lists had the constraint that the same voice could be presented at most in three consecutive trials, while the same sentence could be presented at most in two consecutive trials.

Voice Discrimination Task (VDT)

The auditory stimuli employed in the VDT corresponded to those composing the Montreal Affective Voices set (Belin et al., 2008). Ten different speakers (5 female) recorded nine affective interjections using the vowel /ɑ/ (mean duration = 1000 ms; SD = 540 ms). The VDT followed an AX discrimination design. A trial commenced with the presentation of a fixation point for 1 second. Subsequently, two auditory stimuli were sequentially presented with an ISI of 1 second. Participants had to indicate via button press as fast as they could if the same or different speakers had enunciated the two stimuli without the pressure of a time limit. 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. The two stimuli employed in each trial were always from speakers of the same sex and always corresponded to different affective categories to avoid repeating the same stimuli should a single speaker be employed for both vocalizations. Fifty-two trials (13 trials x 2 sexes x 2 correct answer possibilities) composed the VDT. Trial lists were pseudo-randomly generated to ensure that speakers were never repeated in consecutive trials.

Categorization Task (CT)

The CT was similar to the task first introduced by Pallier and collaborators (1997). The stimuli employed in the CT consisted of a continuum of seven synthesized vowel stimuli between the Catalan vowels /e/and /ε/ with a duration of 185ms. The stimuli were modelled in accordance with a prototypical male voice, i.e., F0 = 120Hz (Mathias & von Kriegstein, 2019). The vowels differed only in the frequency of the first formant that increased from 404 Hz (s1) to 584 Hz (s7) in steps of 30 Hz. For all stimuli, F2 was set to 2084 Hz. The CT consisted of presenting stimuli from the continuum, which participants had to categorize as an example of either /e/ or /ε/. A trial commenced with the presentation of a fixation point. After 1 second, a stimulus was presented while the fixation point remained on display. Participants had to respond 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* (/pεrə/, which means pear) as fast as they could without an imposed time limit. The CT was composed of a total of 63 trials (7 stimuli x 9 repetitions of each). Trial lists were pseudo-randomly generated to avoid the presentation of a stimulus in consecutive trials.

Lexical Decision Task (LDT)

The LDT employed in this study was similar to the lexical decision task described in a previous study (Sebastian-Galles et al., 2005). The stimuli consisted of 344 auditory stimuli (experimental and control) enunciated by a male native Catalan speaker (mean duration = 655 ms; SD = 125 ms). The experimental stimuli included 132 words that contained one of the two phonemes from the targeted Catalan contrast (i.e., /e/ or /ε/) and 132 non-words. Non-words were designed by substituting the /e/ and /ε/ present in the real words with the other member of the phoneme pair. For example, the non-word /mel/ was created from the real word /mεl/, which means honey. Furthermore, 80 control stimuli, 40 words, and 40 non-words were also employed. The control non-words were created by changing a vowel phoneme in a real Catalan word with a phoneme employed in Spanish and Catalan. For instance, based on the real Catalan word /kɔp/ (“hit” in Catalan), the non-word /kup/ was created. The real Catalan words from which the control non-words were created were not employed as control words. The LDT consisted of the presentation of words and non-words, which participants had to classify as part of the Catalan lexicon or not. Participants were informed that non-words would be similar to real words but with mispronounced vowels. A trial began with the presentation of a fixation point during 500ms, after which participants were presented with a stimulus while the fixation point remained on display. Participants had to respond via button press if the auditory stimulus they had heard corresponded to a Catalan word or not. After they had delivered their response and a 750ms period had elapsed, the following trial commenced. If no answer was delivered 5 seconds after the offset of a stimulus, the subsequent trial would commence. The experimental words were distributed between two lists to ensure that participants only heard one member of the same word pair. Lists were counterbalanced across participants. Furthermore, all participants were presented with all control stimuli. Therefore, the LDT was composed of 212 trials (66 experimental words trials + 66 experimental non-words trials + 80 control stimuli trials).

Non-Word Association Task (NWAT)

The Non-Word Association Task was initially introduced in a previous study (Díaz et al., 2022). The visual stimuli consisted of six avatars created with the free-to-use website ([www.crearunavatar.com](http://www.crearunavatar.com)) and differed from those employed in the VRTs. Furthermore, six non-words enunciated by a single native Spanish speaker constituted the audio stimuli for this task (mean duration = 570 ms; SD = 0 ms). The NWAT sought to train and test participants' ability to learn audiovisual associations. It was composed of two phases: a training and a test. The training phase of the NWAT sought to teach the participants associations between non-words and avatars. During the training, a trial commenced with the presentation of a fixation point for 1 second, after which a non-word and its corresponding avatar were simultaneously presented. Each non-word and avatar association was presented twice. Therefore, the training phase comprised 12 trials (6 audiovisual associations x 2 repetitions of each association). Upon completing the training phase, the test phase commenced. The training phase and the test phase employed the same stimuli. A test trial consisted of the display of a fixation point for 1 second, followed by the presentation of a non-word, while the six avatars were presented with a number displayed below them (1 - 6). Participants indicated via button press which avatar they believed was associated with the non-word they had heard as fast as possible without the pressure of a time limit. Eight stimuli lists were created with the constraint that the same non-word could be presented, at most, in two consecutive trials. Each list had different non-word-avatar pairings a total of 48 trials (8 repetitions of each non-word – avatar pairing).

References:

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

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*, 1-9. https://doi.org/10.3389/fpsyg.2022.1008963

Mathias, S. R., & von Kriegstein, K. (2019). Voice Processing and Voice-Identity Recognition. En K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (Vol. 69, pp. 175-209). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4\_7

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

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

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