DOES VISUOSPATIAL WORKING MEMORY PREDICT L2 PERCEPTUAL LEARNING FROM PHONETIC TRAINING WITH HAND GESTURES?

Xiaotong Xi[1], Peng Li[2], Pilar Prieto[3,1]

[1]Universitat Pompeu Fabra, [2]University of Oslo, [3]Catalan Institution for Research and Advanced Studies
xiaotong.xi@upf.edu, peng.li@iln.uio.no, pilar.prieto@upf.edu

ABSTRACT

This study assessed whether visuospatial working memory (VSWM) can predict L2 perceptual learning through audiovisual phonetic training with or without hand gestures. Ninety-nine Catalan speakers were trained on the perception of English vowel pairs /æ-ʌ/ and /i-ɪ/ under one of the following three conditions: training without gestures, with hand gestures cueing lip shape, or with hand gestures cueing tongue position. We assessed participants' VSWM via a symmetry span task and vowel perception accuracy through a word identification task before and after training. Results showed that VSWM positively predicted the perceptual learning of /i-ɪ/ only in the no gesture condition, while no such relationship was found in the two gesture conditions or in the learning of /æ-ʌ/. This suggests that VSWM resources might be recruited during audiovisual phonetic training for processing of subtle visual articulation (e.g., the lip spreading of /i-ɪ/ in the no gesture condition) rather than visually salient articulation (e.g., the lip aperture of /æ-ʌ/ in the no gesture condition) or hand gestures movements.

Keywords: Hand gestures, L2 perception, VSWM, phonetic training, English vowel

1. INTRODUCTION

While phonetic training has shown benefits in L2 perceptual phonological learning [1], recent studies showed that training L2 pronunciation with multimodal input was more effective. On top of auditory input, audiovisual training that provides visual input of articulatory information could facilitate L2 pronunciation more than auditory training (e.g., [2]). Moreover, providing kinesthetic information such as hand gestures in audiovisual training showed further benefits for the non-native speech sound production [3]–[5]. However, less strong effects have been found on the perceptual learning of difficult L2 contrasts. For example, hand gestures encoding durational features could help improve the perception accuracy of L2 Japanese long and short vowels, but the effects were not greater than training without hand gestures (e.g., [4], [6]). Similar findings were also obtained when using gestures encoding aspiration feature to train L2 Mandarin aspirated plosives [3], [7].

It might be that individual differences accounted for the null results found in gestural training on L2 speech perception. In terms of auditory training, individual differences such as perceptual ability, phonological awareness [8], auditory selective attention, attention switching [9], and phonological working memory [10] were shown to predict learning outcomes. However, few studies have investigated whether individual differences in visuospatial processing abilities could predict the effectiveness of gestures in L2 phonetic training.

While working memory has been considered an important factor for various aspects of L2 development [11], the visuospatial sketchpad which stores visual and spatial information [12] may be crucial for learners to process multimodal information during learning. A recent study showed that learners with high visuospatial working memory (VSWM) improved more in math through training with hand gestures, whereas no such effect was found for training without hand gestures [13]. Nevertheless, to the best of our knowledge, no previous study has looked into whether VSWM could predict the learning outcome of gestural training in L2.

This study thus investigates the role of VSWM in the perceptual learning of difficult L2 vowel contrasts through audiovisual phonetic training with or without gestures. Two pairs of English vowels /æ-ʌ/ and /i-ɪ/ were chosen. Both pairs are challenging for Catalan learners of English, as they tend to perceive /æ/ and /ʌ/ as Catalan /a/ [14] and rely more on durational cues in distinguishing /i/ and /ɪ/ than native English speakers [15]. In addition, the vowels also differ in articulation. Specifically, /æ/ is produced with a larger lip aperture and more fronted tongue position than /ʌ/ [16], and /i/ shows wider lip spreading and higher tongue position than /ɪ/ [17]. Two types of hand gestures were designed to highlight these key articulation differences for each vowel pair (Figure 1). The "lip hand gesture" illustrated the differences in lip aperture for /æ-ʌ/ and lip spreading for /i-ɪ/ by the distance between the thumb and the four fingers; the "tongue hand gesture" mirrored the mouth roof by one hand and the tongue movement by the other hand to show the tongue positions for /æ-ʌ/ and /i-ɪ/.

Based on previous studies [13], we hypothesized that VSWM abilities would positively predict the perceptual learning of L2 sound contrasts in the phonetic training with hand gestures, while it does not predict the learning outcome with no gestures.

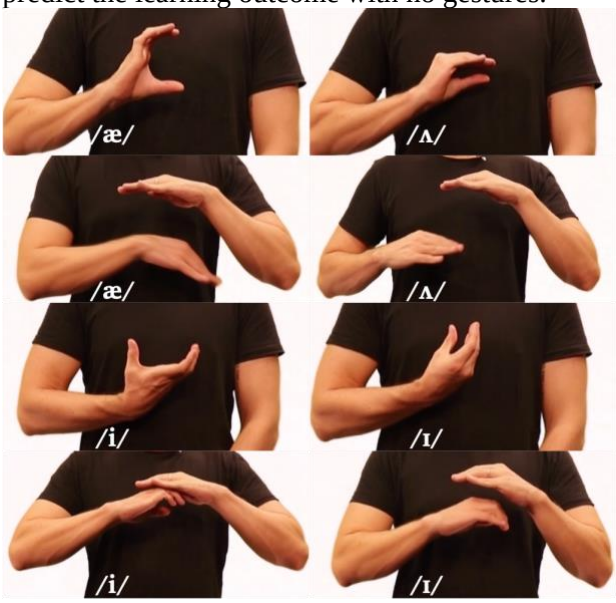

Figure 1: From top to bottom: lip hand gestures for /æ/-/ʌ/, tongue hand gestures for /æ/-/ʌ/, lip hand gestures for /i/-/ɪ/, and tongue hand gestures for /i/-/ɪ/.

2. METHOD

2.1. Participants

Ninety-nine Catalan/Spanish bilingual learners of English (female = 84, M_{age} = 19.7, *SD* = 1.8) were recruited from a public university in Barcelona. They reported having an intermediate English level. They were randomly assigned to one of the three training conditions, with each condition having 33 participants: No Gesture (NG, 26 females), Lip Hand Gesture (LG, 28 females), and Tongue Hand Gesture (TG, 30 females). Participants within each condition received two training sessions, one for /æ-ʌ/ and the other one for /i-ɪ/, in counterbalanced order.

2.2. Materials

2.1.1. Phonetic training

An American English male speaker was video recorded in a broadcasting studio while producing the materials of the two vowel pairs (/æ-ʌ/ and /i-ɪ/) for the three conditions (i.e., NG, LG, and TG). Each training session consisted of a familiarization phase and a training phase. For the familiarization phase, the instructor introduced the vowel pairs explaining the articulatory differences. In addition, the instructor presented the hand gestures and explicitly explained what they represented for the two gesture conditions.

For the training phase, 6 minimal pairs of English CVC words were selected for each vowel pair (e.g., *cat-cut* /kæt-kʌt/, *beat-bit* /bit-bɪt/). Then, for each training word, a short sentence was created using the word (e.g., A CAT walks by). For the NG condition, the instructor produced the training stimuli without any hand movements. For the LG and TG conditions, the instructor performed the corresponding lip or tongue hand gestures while producing target vowels.

In the end, the familiarization and training videos were uploaded to the Tobii Pro Lab software to create six training videos (2 vowel pairs × 3 conditions). In each video, the familiarization video preceded the training video clips which contained 3 repetitions of training words and sentences. In total, the duration of each training video was about 15 minutes.

2.1.2. Perception test: identification task

A word identification task was used to measure participants' perception accuracy of the two vowel contrasts before, right after training, and one week after training. For each vowel contrast, a total of 18 CVC words were selected. The instructor recorded a total of 36 words and the audio files were uploaded to Alchemer (www.alchemer.com) to build 2 surveys, one for each vowel contrast.

2.1.3. VSWM

We created a symmetry span task adapted from [18] on the platform Pavlovia (https://pavlovia.org) to test participants' VSWM capacity. The task consisted of the spatial recall task and the symmetry judgment task. In the spatial recall trials, a 4 × 4 matrix with a square filled with red was presented. In the follow-up symmetry judgment trial, another 4 × 4 matrix was presented with some squares filled in black and two response options of 'yes' and 'no' below. After 2 to 6 sequences of spatial recall and symmetry judgment, a 4 × 4 matrix was presented. Figure 2 exemplifies a trial of 2 sequences.

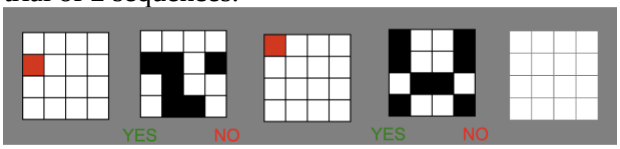

Figure 2: Example symmetry span trial of 2 sequences.

2.3. Experimental procedure

All participants signed a consent form prior to the experiment. They received the training and tests individually in a soundproof room. Half of the participants received the /æ-ʌ/ training first, the other half were trained on /i-ɪ/ first. Before each training session, participants were assessed on their perception of the target vowel pair through the word

identification task. They listened to each of the testing words and had to choose the correct answer from one of the two options which were minimal pair words contrasting only in /æ-ʌ/ or /i-ɪ/. Participants then watched training videos. The LG and TG group saw the instructor produce the gestures accompanying the speech, whereas the NG group only saw the instructor produce the same speech. Participants performed the identification task immediately after the training session. At the end of the experiment, participants completed the VSWM task in which they were instructed to recall the red squares in the correct location and serial order and to decide whether the black-square design was symmetrical or not in between. One week later, participants performed the delayed posttest by using the same identification task. Note that the testing words of the identification task were identical across all three tests, with the order being automatically randomized by the platform.

2.4. Data coding and statistical analyses

For the identification task, participants' correct responses scored 1, and incorrect responses scored 0. The pretest, posttest, and delayed posttest scores of /æ-ʌ/ and /i-ɪ/ were exported from Alchemer. Then, the improvement scores of each vowel pair were calculated for each participant. Specifically, the immediate improvement score was obtained by subtracting the identification sum score of the posttest from that of the pretest and the sustained improvement score was calculated by subtracting the identification sum score of the delayed posttest from that of the pretest. The VSWM score was calculated for each participant by summing up the number that they recalled the position of the red squares in the correct order [18]. One participant from the TG condition scored 0 due to a misunderstanding of the instruction. Thus, the VSWM score of this participant was removed from the database.

In order to check whether participants' VSWM capacity could predict L2 perceptual learning through phonetic training, we ran two Linear Mixed Models (LMMs) using the *lme4* package [19] for the immediate improvement score and sustained improvement score. The fixed effects included Condition (3 levels: NG, LG, and TG), Vowel Pair (2 levels: /æ-ʌ/ and /i-ɪ/), VSWM score, and their interactions. Random factors included a by-participant intercept. The significance for the fixed effects was calculated with Type II Wald Chi-squared tests using the *car* package [20]. For the post-hoc analysis, when significant interactions were observed with VSWM, we examined the slopes of VSWM and conducted pairwise comparisons with Bonferroni adjustment using *emtrends* function from the *emmeans* package [21].

3. RESULTS

Table 1 shows the descriptive data of the perceptual improvement scores across groups and vowel pairs.

	/æ-ʌ/	/i-ɪ/
Posttest – Pretest		
NG	0.1 (1.2)	2.0 (2.2)
LG	0.2 (1.6)	2.4 (2.4)
TG	0.4 (1.8)	2.2 (2.3)
Delayed posttest - Pretest		
NG	-0.3 (2.1)	1.6 (2.2)
LG	0 (1.7)	2.1 (2.2)
TG	0.1 (1.6)	1.5 (2.0)

Table 1: Mean (SD) of perceptual improvement scores across three groups and vowel pairs.

3.1. Improvement from pretest to posttest

Results of the LMM for the immediate improvement revealed a significant main effect of vowel pair (χ^2 = 136.3, $p < .001$) and a significant 3-way interaction of Condition × Vowel Pair × VSWM (χ^2 = 12.9, p = .002). Figure 3 plotted the model prediction on the immediate perceptual improvement from the 3-way interaction. The post-hoc results revealed a significantly positive slope of VSWM in the NG condition when learning /i-ɪ/ (β = 0.06, 95% CI = [0.01, 0.11], p = .028), suggesting a positive relationship between VSWM and L2 perceptual learning of /i-ɪ/ when gesture was absent. Pairwise comparisons showed that the positive slope of VSWM in the NG for the learning of /i-ɪ/ significantly differed from that in the LG (Δ = 0.11, p = .043).

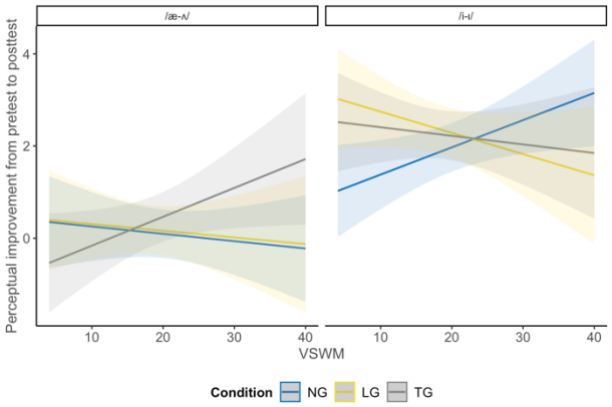

Figure 3. LMM prediction of the perceptual improvement from pretest to posttest across condition, vowel pair, and VSWM. The bands represent the 95% CI.

3.2. Improvement from pretest to delayed posttest

Results of the LMM for the sustained improvement revealed a significant main effect of vowel pair (χ^2 = 89.0, p < .001), a significant 2-way interaction of Condition × VSWM (χ^2 = 6.8, p = .033), and a significant 3-way interaction of Condition × Vowel Pair × VSWM (χ^2 = 7.3, p = .026). Figure 4 showed the model prediction on the sustained improvement from the 3-way interaction. The post-hoc analysis of the 3-way interaction showed that VSWM was a significant predictor of perceptual learning of /i-ɪ/ in the NG condition (β = 0.08, 95% CI = [0.02, 0.13], p = .008). These results suggest that better VSWM abilities predict better sustained improvement in the perception of /i-ɪ/ contrast in the NG condition. Pairwise comparisons revealed that the positive slope observed in the NG condition was significantly higher than those in the LG (Δ = 0.15, p = .006) and TG conditions (Δ = 0.12, p = .017).

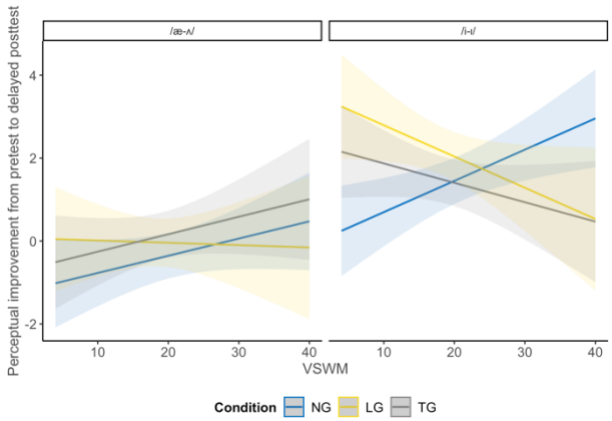

Figure 4. LMM prediction of the perceptual improvement from pretest to delayed posttest across condition, vowel pair, and VSWM. The bands represent the 95% CI.

4. DISCUSSION

This study examined whether VSWM predicts the perceptual learning of L2 English vowels /æ-ʌ/ and /i-ɪ/ through audiovisual phonetic training with or without hand gestures cueing articulation. We assessed the perceptual learning outcome through an identification task and learners' VSWM via a symmetry span task. The hand gestures cued either the lip shape or the tongue movement information. We found that VSWM predicted the learning outcome of /i-ɪ/ in the no gesture condition.

First, contradictory to our hypothesis, VSWM could not significantly predict the identification accuracy of the pair of /i-ɪ/ sounds in audiovisual phonetic training with either of the two types of hand gestures. This result is inconsistent with the results in [13], which might be due to the gesture type and function. Our gestures iconically represented the articulatory movements of the speech sounds but the pointing gestures in [13] pointed to the learning target *per se*. Therefore, if students were good at processing spatial information, gestures would help the learning process by pointing to the target. In our case, the learning target contained abstract knowledge. The processing of this information might require cognitive abilities beyond visuospatial processing abilities. Therefore, VSWM was not a significant predictor of the gesture conditions.

Second, VSWM was not a significant predictor of the learning of /æ-ʌ/ in any of the conditions, including the no gesture condition. A pilot study with 15 participants found that native English speakers distinguished /æ-ʌ/ better than /i-ɪ/ (85% vs. 60%) through visual-only input. Therefore, we hypothesized that the visually subtle articulation differences between /i/ and /ɪ/ would require more VSWM resources. Another possible explanation could be due to the difference in improvement between the two vowel pairs. As noted in Table 1, learners' perception accuracy of /i-ɪ/ improved more than /æ-ʌ/. With such a small improvement and variation (indicated by mean and SD) in the learning of /æ-ʌ/, it could be hard to observe a clear correlation between improvement and VSWM.

This is an initial attempt to assess the role of VSWM in learning L2 speech sounds through hand gestures. There might be several factors that can affect the results. For example, the way in which we test the VSWM (e.g., spatial span task, Corsi's block-tapping task, etc.), the measure for assessing the learning (e.g., perceptual measure or productive measure, subject rating or acoustic analysis, etc.).

In conclusion, this study shows that VSWM can predict perceptual learning of L2 sounds in situations where the visual information is subtle, e.g., in the no gesture conditions where the visual difference of /i-ɪ/ lies in the subtle variation in lip spreading. VSWM resources may thus be recruited in more challenging phonetic training tasks but might not be needed in the processing of visually salient hand gestures cueing articulation during phonetic training.

5. ACKNOWLEDGMENTS

This study was supported by the Ministry of Science, Innovation, and Universities, State Research Agency and European Regional Development Fund [PGC2018-097007-B-I00], and by the Ministry of Science and Innovation [PID2021-123823NB-I00]. The first author is supported by the Ministry of Research and Universities of Catalonia and the European Social Fund [2022FI_B2 00209]. The second author is supported by the Research Council of Norway [223265]. We thank Patrick Louis Rohrer for helping with the creation of materials.

6. REFERENCES

[1] R. I. Thomson, "High Variability [Pronunciation] Training (HVPT)," *J. Second Lang. Pronunciation*, vol. 4, no. 2, pp. 208–231, Dec. 2018, doi: 10.1075/jslp.17038.tho.

[2] V. Hazan, A. Sennema, M. Iba, and A. Faulkner, "Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English," *Speech Commun.*, vol. 47, no. 3, pp. 360–378, 2005, doi: 10.1016/j.specom.2005.04.007.

[3] X. Xi, P. Li, F. Baills, and P. Prieto, "Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features," *J. Speech, Lang. Hear. Res.*, vol. 63, no. 11, pp. 3571–3585, 2020, doi: 10.1044/2020_JSLHR-20-00084.

[4] P. Li, F. Baills, and P. Prieto, "Observing and Producing Durational Hand Gestures Facilitates the Pronunciation of Novel Vowel Length Contrasts," *Stud. Second Lang. Acquis.*, vol. 42, no. 5, pp. 1015–1039, 2020, doi: 10.1017/S0272263120000054.

[5] M. Hoetjes and L. van Maastricht, "Using gesture to facilitate L2 phoneme acquisition: the importance of gesture and phoneme complexity," *Front. Psychol.*, vol. 11, p. 575032, 2020, doi: 10.3389/fpsyg.2020.575032.

[6] Y. Hirata and S. D. Kelly, "Effects of lips and hands on auditory learning of second-language speech sounds," *J. Speech, Lang. Hear. Res.*, vol. 53, no. 2, pp. 298–310, 2010, doi: 10.1044/1092-4388(2009/08-0243).

[7] P. Li, X. Xi, F. Baills, and P. Prieto, "Training non-native aspirated plosives with hand gestures: learners' gesture performance matters," *Lang. Cogn. Neurosci.*, vol. 36, no. 10, pp. 1313–1328, 2021, doi: 10.1080/23273798.2021.1937663.

[8] T. K. Perrachione, J. Lee, L. Y. Y. Ha, and P. C. M. Wong, "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," *J. Acoust. Soc. Am.*, vol. 130, no. 1, pp. 461–472, 2011, doi: 10.1121/1.3593366.

[9] I. Mora-Plaza, M. Ortega, and J. C. Mora, "High-Variability Phonetic Training Under Different Conditions: Individual Differences in Auditory Attention Control," in *Theoretical and Practical Developments in English Speech Assessment, Research, and Training*, V. G. Sardegna and A. Jarosz, Eds. Springer Nature Switzerland AG, 2022, pp. 241–260.

[10] C. Aliaga-García, J. C. Mora, and E. Cerviño-Povedano, "L2 speech learning in adulthood and phonological short-term memory," *Poznań Stud. Contemp. Linguist.*, vol. 47, no. 1, pp. 1–14, Jan. 2011, doi: 10.2478/psicl-2011-0002.

[11] J. A. Linck, P. Osthus, J. T. Koeth, and M. F. Bunting, "Working memory and second language comprehension and production: A meta-analysis," *Psychon. Bull. Rev.*, vol. 21, no. 4, pp. 861–883, 2014, doi: 10.3758/s13423-013-0565-2.

[12] A. Baddeley, "Working memory: Theories, models, and controversies," *Annu. Rev. Psychol.*, vol. 63, pp. 1–29, 2012, doi: 10.1146/annurev-psych-120710-100422.

[13] M. Aldugom, K. Fenn, and S. W. Cook, "Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity," *Cogn. Res. Princ. Implic.*, vol. 5, p. 27, 2020, doi: 10.1186/s41235-020-00215-8.

[14] J. Cebrian, "Perception of English and Catalan vowels by English and Catalan listeners: A study of reciprocal cross-linguistic similarity," *J. Acoust. Soc. Am.*, vol. 149, no. 4, pp. 2671–2685, 2021, doi: 10.1121/10.0004257.

[15] J. Cebrian, "Experience and the use of non-native duration in L2 vowel categorization," *J. Phon.*, vol. 34, no. 3, pp. 372–387, 2006, doi: 10.1016/j.wocn.2005.08.003.

[16] J. P. Zerling, "Frontal lip shape for French and English vowels," *J. Phon.*, vol. 20, no. 1, pp. 3–14, 1992, doi: 10.1016/s0095-4470(19)30249-9.

[17] S. A. J. Wood, "A radiographic and model study of the tense-lax contrast in vowels," in *Phonologica 1988*, 1992, pp. 283–291.

[18] K. J. Blacker, S. M. Weisberg, N. S. Newcombe, and S. M. Courtney, "Keeping track of where we are: Spatial working memory in navigation," *Vis. cogn.*, vol. 25, no. 7–8, pp. 691–702, 2017, doi: 10.1080/13506285.2017.1322652.

[19] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015, doi: 10.18637/jss.v067.i01.

[20] J. Fox and S. Weisberg, *An {R} Companion to Applied Regression*. SAGE, 2019.

[21] R. V Lenth, "emmeans: Estimated Marginal Means, aka Least-Squares Means." 2022.