Embodied speech: sensorimotor contributions to native and non-native language processing and learning

Tzuyi Tseng, Jennifer Krzonowski, Claudio Brozzoli, Alice Catherine Roy, Véronique Boulenger

Embodied speech: sensorimotor contributions to native and non-native language processing and learning

Tzuyi Tseng[1], Jennifer Krzonowski[1], Claudio Brozzoli[2]*, Alice C. Roy[1]*, Véronique Boulenger[1]*

[1] Dynamique du Langage, UMR 5596 CNRS - Université Lumière Lyon 2, Lyon, France

[2] ImpAct team, Centre de Recherche en Neurosciences de Lyon, INSERM, Lyon, France

* co-last authors

Keywords: Embodiment, Speech Perception, Motor System, Non-Native Phonemes, Foreign Language Learning, Manual Gestures

Abstract

The last 15 years of research on language production and comprehension have put forward a conceptual revolution, highlighting how language may depend not only on specific brain areas but also on embodied processes underpinned by sensorimotor brain regions. A large amount of studies have underlined the sensorimotor grounding of the different processes at play in language comprehension, in particular action verb processing and sentence comprehension (see Fischer & Zwaan, 2008 and Franken et al., 2022 for reviews). One aspect of language processing, phoneme perception, has been less extensively studied in light of embodied cognition, yet the motor involvement in decoding sounds pertaining to a linguistic code seems to play a crucial contribution. The focus of the current review is to provide an updated overview of the findings in this domain on both native (*Sections I and II)* and foreign (i.e. non-native) phoneme perception (*Section III)*, and to further examine if and how motor training may enhance phonological processing and learning of foreign language speech sounds (*Section IV and V*).

I. Motor resonance to native speech perception

Neuroimaging studies using Transcranial Magnetic Stimulation (TMS) or functional Magnetic Resonance Imaging (fMRI) have provided compelling evidence that (pre)motor regions involved in speech production are also activated during the mere perception of native phonemes. In their seminal work, Fadiga and colleagues (2002) applied TMS to the left motor cortex of Italian adults while they listened to words and pseudo-words featuring either the double lingua-palatal fricative (alveolar trill) /rr/, which requires strong tongue tip movements to be produced, or the double labiodental fricative /ff/, which only involves minor tongue

movements. Items embedded with /rr/ evoked higher tongue motor evoked potentials (MEPs) than the other conditions, showing automatic and somatotopic activity in the motor cortex for passive speech perception. Roy and colleagues (2008) replicated this phonological effect with Italian pseudo-words including the tongue-related /ll/, and further revealed an early lexical effect by comparing motor resonance to this double consonant embedded in rare or frequent words. When TMS was applied 200 and 300 ms after the double consonant, the mere perception of /ll/ in rare Italian words evoked larger MEPs than /ll/ in frequent words. This finding suggests that the motor cortex does not only participate in phonological encoding during speech perception, but also interacts with top-down lexical processes. Additional evidence for speech-induced motor activity came from D'Ausilio and colleagues (2014) who applied ultrasound tissue Doppler imaging (UTDI) to record the whole tongue movement synergies evoked by TMS during speech perception. Passively listening to syllables varying in place of articulation and position in the vowel space (/ti/, /to/, /ki/ and /ko/) indeed mirrored the kinematic patterns of actual phoneme production, with tongue displacement along the anterior-posterior (for coronal /t/ and velar /k/, respectively) and ventral-dorsal (for high-front /i/ and back /o/ vowels, respectively) planes. In line with these findings, studies using fMRI also revealed a partial overlap of motor cortical activation during speech perception and production. Wilson and colleagues (2004) pioneered in showing stronger hemodynamic response in the bilateral ventral premotor cortex, activated for speech production, when merely perceiving consonant-vowel (CV) syllables compared to nonspeech sounds (white noise or bell rings). Pulvermüller and colleagues (2006) furthermore provided evidence for the specificity of this motor resonance, as consonants requiring different places of articulation for production activated the precentral gyrus somatotopically. Whereas perception of the labial /p/ induced activity in the lip motor area, listening to the dental /t/ increased activity in the tongue motor area. This finding of distinct representations for tongue- and lip-related speech sounds in the precentral gyrus underlies the embodiment of phonemes. Other fMRI studies however failed to replicate this somatotopic encoding of phonemes in the (pre)motor cortex during passive speech perception. Arsenault and Buchsbaum (2016) for instance reported equal activity of the lip and tongue motor representations when listening to labial and dental/alveolar consonants. This was confirmed by additional Multivariate Pattern Analyses (MVPA): a classifier trained to discriminate these same consonants from production data in the (pre)motor cortex (i.e. pre/postcentral gyri and central sulcus) failed to discriminate them when they were only perceived (see also Cheung et al., 2016 for somatotopy for place of articulation in speech production but not perception). Despite these discrepant findings, evidence still shows that articulatory features of phonemes can be decoded from cortical activity in motor and premotor regions. Using a similar MVPA approach, Correia and colleagues (2015) trained a classifier to discriminate between syllable pairs based on either place of articulation, manner of articulation or voicing. They then tested whether the classifier could predict brain activity from other phonemes varying on the same articulatory properties. For instance, they trained a classifier to discriminate place of articulation between labials and

dentals in plosives (/pa/ vs /ta/) and tested its ability to discriminate this same feature in fricatives (/fa/ vs /sa/). Similarly, discrimination of manner of articulation across two places of articulation was tested (i.e. training on labial plosives vs fricatives: /pa/ vs /fa/, and test on dental/alveolar plosives vs fricatives: /ta/ vs /sa/). Results revealed that the classifier generalized to discriminate both place and manner of articulation from activity patterns in a network distributed over the bilateral postcentral gyrus and the right anterior insula. Generalization to place of articulation further extended to the bilateral superior temporal cortex, as well as the precentral and inferior frontal gyri in the right hemisphere. The temporo-parietal junction also decoded these two articulatory features, in the left hemisphere for place and in the right for manner of articulation (see also Archila-Meléndez et al., 2018 for converging results of generalization in place of articulation). On the other hand, successful classification of voicing (i.e. training on unvoiced vs voiced plosives: /p/ vs /b/, and test on unvoiced and voiced fricatives: /f/ vs /v/) specifically activated the right anterior superior temporal sulcus. Overall, these studies therefore support that in addition to temporo-parietal and inferior frontal regions, premotor and motor areas code for articulatory features even when phonemes are only auditorily perceived, highlighting the sensorimotor nature of speech perception (Hickok & Poeppel, 2004, 2007; Schwartz et al., 2012).

II. Functional contribution of the motor cortex in speech perception

The question of the role of the motor system as an essential or a subsidiary component of speech perception has been hotly debated (Hickok, 2011; Lotto et al., 2009; Pulvermüller & Fadiga, 2010; Schomers & Pulvermüller, 2016; Scott et al., 2009; Skipper et al., 2017). Evidence for tight reciprocal functional links between perception and production first comes from behavioral studies. Using electropalatography in healthy adults, Yuen and coworkers (2010) reported that perceiving incongruent distractor speech sounds during syllable production distorted the ongoing articulatory gestures. For instance, pronouncing /ka/ or /sa/ induced closer contact of the tongue against the alveolar ridge while hearing /ta/ rather than the same congruent syllables. This suggests that phoneme perception automatically activates articulatory movements, thus interfering with actual production. Reciprocally, changes in the articulatory configuration can impact speech perception in both adults (Ito et al., 2009) and infants (Bruderer et al., 2015; Choi et al., 2019). In 6 months-old infants, temporarily restraining either the tongue tip or the closure of the lips with teething toys indeed impaired the discrimination of non-native dental and labial speech sounds, respectively. This supports an active contribution of motor processes in speech perception early in the course of language development. TMS studies also provided strong evidence for a causal motor role in speech perception. D'Ausilio and colleagues (2009) revealed faster recognition of labials (e.g. /p/) masked by white noise when activity of the lip motor area was primed with TMS, but faster recognition of dentals (e.g. /t/) for stimulation of the tongue

motor area. Conversely, temporarily disrupting left premotor cortex activity by repetitive TMS (rTMS) altered the discrimination of plosives in noisy CV syllables (/pa/, /ta/, /ka/) (Meister et al., 2007; see also Murakami et al., 2015 and Sato et al., 2009 for converging findings). Möttönen and Watkins (2009) furthermore showed that this effect is specific to phonemes' articulatory features as well as to the targeted motor area. rTMS over the left lip motor representation impaired categorical perception of a place-of-articulation continuum ranging between lip- and tongue-related phonemes (/ba/-/da/). This effect was not seen for a voice-onset-time continuum including phonemes that are not lip-articulated (/ka/-/ga/), nor when the left-hand motor cortical representation was disrupted with rTMS.

What is particularly striking from some of the studies reviewed so far is that motor regions seem to be preferentially engaged when speech perception is challenging. As a matter of fact, in a follow-up of their TMS study conducted in 2009, D'Ausilio and colleagues (2012) reported facilitation of dental and labial consonant identification by, respectively, tongue and lip motor area stimulation for noise-embedded but not intact syllables. Converging evidence for specific motor activity to degraded speech comes from Callan and coworkers (2010) who highlighted the contribution of the ventral premotor cortex in distinguishing between correct and incorrect phoneme identification in noise. Similarly, using a continuum ranging from white noise to CV syllables, Osnes and collaborators (2011) found an increase of fMRI hemodynamic activity in superior temporal regions, with a left-hemisphere bias, as sounds became progressively more recognized as speech. Crucially, the left premotor cortex was specifically activated when sounds became identifiable as speech but were still noisy, whereas temporal cortical activity ceased to increase in this condition. Effective connectivity analyses in the left hemisphere further showed bidirectional connections between the premotor cortex and superior temporal sulcus, together with unidirectional transfer from the planum temporale to the premotor cortex (see also Alho et al., 2014). These results first confirm that premotor regions selectively come into play when processing degraded though identifiable speech sounds. They also concur with the view that (pre)motor regions are part of a sensorimotor circuit that maps articulatory and auditory representations through the use of internal models, which may constrain and facilitate speech perception especially in challenging listening situations (see Callan et al., 2004; Rauschecker & Scott, 2009; Skipper et al., 2007; Wilson et al., 2004). Du and colleagues (2014) reached similar conclusions by showing increasing left ventral premotor cortex activity as a function of decreasing signal-to-noise ratio (SNR), namely when noise-embedded phonemes were less accurately identified. Temporal cortex activity on the contrary positively correlated with behavioral performance. Importantly, multivariate analyses (MVPA) revealed that premotor regions successfully categorized phonemes at moderate and high SNRs (SNR ≥ -6), whereas temporal regions only exhibited good phoneme categorization in the absence of noise. Premotor regions therefore appear more robust to noise than auditory regions, suggesting they can help speech processing, in moderate noisy conditions at least.

Motor activity has also been reported for other types of speech degradation, from noise-vocoding (Hervais-Adelman et al., 2012) to speech rate acceleration (Adank & Devlin, 2010; Hincapié Casas et al., 2021) and inter-talker variability (Bartoli et al., 2015). In a series of TMS studies, Nuttall and collaborators (2016, 2017) actually demonstrated that this motor resonance occurs irrespective of the nature of the degradation, either extrinsic or intrinsic to the speech signal. TMS-induced lip MEPs were significantly larger during perception of distorted than of intact speech sounds, whether the distortion resulted from white noise masking or from obstructing lip and tongue movements. Nuttall and colleagues furthermore observed larger lip motor activity for labials (/aba/, /apa/) than for dentals (/ada/, /ata/) but only in the distorted conditions (in line with D'Ausilio et al., 2012). Two other findings were of primary interest. First, participants who were better at identifying the degraded syllables showed larger lip MEPs during passive perception, compared to low performers (Nuttall et al., 2016). In other words, stronger motor activity to speech was associated with better recognition of degraded speech (see D'Ausilio et al., 2014 for converging results). Second, participants' hearing sensitivity influenced motor recruitment during perception (Nuttall et al., 2017). Whereas speech motor facilitation was found for noisy speech sounds in participants with better auditory acuity, participants with normal but lower hearing performance showed stronger MEPs for clear speech (see also Du et al., 2016 for similar evidence in younger and older listeners). This suggests that the motor cortex may compensate for impoverished auditory information, resulting either from the signal itself or from a decrease in hearing abilities.

III. Motor resonance to foreign phonemes

Speech motor areas are recruited for the perception of phonemes in the native language, especially under degraded conditions as discussed in the previous section. In parallel, the embodiment of phonemes that are not part of the listener's phonological repertoire has been investigated, although to a much lesser extent. Wilson and Iacoboni (2006) examined the fMRI neural responses in auditory and motor cortices for 25 non-native phonemes (e.g. stops, fricatives, clicks, trills, nasals, etc.) belonging to different languages and varying in producibility for English native speakers. When contrasted to native phonemes, non-native sounds yielded an increased activity in bilateral superior temporal regions. Interestingly, the more difficult the phonemes were judged to produce, the more the temporal cortices were activated. With regard to the motor cortex, only a region-of-interest (ROI) analysis revealed that, for both hemispheres alike, the ventral premotor cortex was more activated for non-native compared to native phonemes (note that whole-brain analyses yet revealed (pre)motor activity for all speech sounds vs rest). In addition, the premotor cortex was found to be functionally connected with superior temporal regions that distinguished non-native from native sounds and that coded for producibility. Wilson and Iacoboni (2006) interpreted their findings in light of internal models instantiated within motor regions to predict the acoustic consequences of the perceived phonemes (see also

Callan et al., 2004). Whereas a match between such predictions and the actual sounds would be rapidly obtained for the native language, the repeated and unsuccessful attempts to simulate unknown, non-native speech sounds would account for the greater motor activity observed. Increased motor cortical activity for non-native phonemes has been corroborated by Schmitz and coworkers (2019) using TMS. The authors probed the lip representation excitability in the left primary motor cortex while Italian participants passively listened to native and non-native German vowels (/a/, /i/, /u/ vs /y/). Echoing Wilson and Iacoboni's (2006) results in the temporal cortices, Schmitz and colleagues (2019) reported a negative correlation between nativeness ratings and the lip motor potentials evoked for vowels: the less the vowel appeared as pertaining to the native repertoire, the higher the excitability in the lip motor representation. The authors suggested a compensatory role of the motor cortex when listening to speech sounds that lack a defined acoustic-motor representation. Such an interpretation fits with the above-reviewed findings on degraded native speech perception (D'Ausilio et al., 2012; Nuttall et al., 2016, 2017) as well as with studies showing stronger left (pre)motor cortical activity for the perception of difficult contrasts in non-native languages (Callan et al., 2003, 2004, 2014). Identification of words starting with the English phonemes /ɹ/ or /l/, that are hardly distinguished by Japanese speakers, has indeed been shown to enhance activity in a bilateral network encompassing articulatory cortical regions in those participants (Callan et al., 2003, 2004). Interestingly, when native speakers of English performed the task on these same English phonemes but produced by Japanese speakers, therefore with a foreign accent, stronger bilateral premotor involvement was also found (Callan et al., 2014).

Altogether, these findings support the idea that listeners make use of brain regions involved in speech production to process speech especially under adverse auditory conditions, be it in their native language that is distorted by noise or mispronounced, or in a foreign language. In this view, although not being strictly essential for speech perception, the motor system seems to play a crucial role in speech sensorimotor integration by constraining phonemic categorization, ultimately facilitating speech perception (Callan et al., 2004, 2014; Iacoboni, 2008; Rauschecker & Scott, 2009; Schwartz et al., 2012; Skipper et al., 2007). Such a functional contribution questions whether learning and processing non-native phonemes could benefit from sensorimotor training. Before presenting recent advances along this line, we will first review the learning paradigms classically developed to support the acquisition of phonemes in a foreign language.

IV. Classical training paradigms for foreign phonemes

Learning speech sounds that are not part of our phonological inventory is challenging, especially in adulthood. Because adult learners cannot rely on robust auditory or articulatory patterns for these newly acquired sounds, they often find them problematic to distinguish from native phonemes. On the production side, this is typically reflected by a native-like pronunciation of the foreign phonemes, a phenomenon commonly

experienced as a foreign accent. The proximity between the phonological systems in the native (L1) and the foreign languages has been advocated as a major factor influencing the learning of the new language's phonemes. According to Flege's Speech Learning Model (Flege, 1995), foreign speech sounds perceived as close to L1 phonemes tend to be assimilated to their native counterparts, and are therefore less well recognized and produced than more distant foreign phonemes. In other words, the greater the perceptual distance between a non-native speech sound and a native phoneme, the more likely and easily it will form a new phonemic category (see also the Perceptual Assimilation Model; Best, 1994 and Best et al., 2001). Despite these difficulties, learning new phonemes has been shown to benefit from laboratory training based on perception and/or production. In this respect, one of the most common training paradigms used to improve foreign speech sound processing is the High Variability Phonetic Training (HVPT), first developed by Logan and colleagues (1991) and embedded in a pre-test/post-test design. HVPT consists in presenting multiple natural tokens of the target phonemes produced by several native speakers in a variety of phonological environments (e.g., varying adjacent phonemes and/or different syllabic positions). Tokens are typically presented from minimal pairs contrasting the native and non-native phonemes, and participants are required to perform a two-alternative forced-choice (2-AFC) identification task with immediate feedback on their response (either on incorrect trials only, or more often on both correct and incorrect trials). Exposing learners to a wide range of acoustic-phonetic cues across different phonological environments during training is thought to enhance perceptual learning and thus to promote the development of new phonemic categories. In addition, providing feedback allows to focus participants' attention on the crucial cues of the speech sounds under consideration (Logan et al., 1991; but see Vlahou et al., 2012 for more robust learning after implicit training without external feedback). Pre- and post-training performance is assessed with the same identification task but without any feedback, and both trained and new tokens, produced by the same or by different speakers not heard during training, are usually included to assess generalization of learning.

Numerous studies have shown improvement of learners' perceptual performance after 3-to-4 weeks of HVPT (from 15 up to 45 training sessions), mostly regarding the English /ɹ/-/l/ contrast that Japanese native speakers typically struggle to discriminate. The benefits of HVPT furthermore generalized to new exemplars and speakers (Bradlow et al., 1997; Callan et al., 2003; Iverson et al., 2005; Lively et al., 1993, 1994; Logan et al., 1991; McClelland et al., 2002; Shinohara & Iverson, 2018), with (moderate) long-term effects up to six months after training (Bradlow et al., 1999; Lively et al., 1994). HVPT can also enhance perceptual performance for other phonological contrasts, such as places of articulation in consonants (Cebrian & Carlet, 2014; Golestani & Zatorre, 2004, 2009; Pruitt et al., 2006), as well as for vowels (Iverson et al., 2012; Nishi & Kewley-Port, 2007, 2008) and tones (Y. Wang et al., 1999, 2003). In a classical HVPT paradigm varying consonantal contexts and speakers (Lambacher et al., 2005), Japanese native speakers for instance exhibited higher identification of vowels from American English at post-test, in particular for those that were more distant from the native

repertoire (/ɔ/ and /ɚ/). Interestingly, perceptual identification training has also proved successful on speech production, despite no explicit articulatory instruction was provided to the learners (see Sakai & Moorman, 2018 for a review). Bradlow and colleagues (1997) reported that the production of words containing /ɹ/ or /l/ by Japanese trainees was rated higher and was better identified by English native speakers after perceptual training than before. In agreement with this study, the American English vowels produced by Japanese learners were better identified by native speakers, and their spectral overlap was reduced at post-test compared to pre-test (Lambacher et al., 2005). This was especially true for more distant vowels (/ɔ/, /ɚ/ and /æ/) whereas vowels (/ɑ/ and /ʌ/) phonetically similar to their Japanese counterpart (/a/) still showed a large degree of overlap after training. These findings support models of second language acquisition (Best, 1994; Flege, 1995) by revealing better learning, both in perception and production, of non-native vowels that share less phonetic features with the native phonological inventory. In addition, they show that transfer of knowledge can occur from perceptual learning to production of non-native phonetic contrasts, highlighting the existence of common auditory-articulatory representations for speech perception and production.

Although HVPT has repeatedly been shown to improve foreign speech sound learning, its effects can vary depending on the learners' native repertoire (e.g. better learning for larger L1 vowel inventory, Iverson & Evans, 2007, 2009), as well as on their perceptual abilities (e.g. detrimental effects for learners with low initial skills, (Perrachione et al., 2011; Sadakata & McQueen, 2014). The source of variability required for efficient learning has also been questioned, especially regarding the use of multiple vs single talkers in HVPT. Whereas the meta-analysis by Zhang and coworkers (2021) found a robust advantage of multi-talker over single-talker training, Brekelmans and colleagues (2022) showed in their review that trainees exposed to high variability in voices did not always outperform those exposed to low variability input. In an attempt to carefully replicate the studies by Logan et al. (1991) and Lively et al. (1993) on the English /ɹ/-/l/ contrast, they found a gain in post-test performance, with generalization to new speakers, both for high variability (including five English native speakers) and low variability (with only one speaker) training, considering learners' initial abilities (see also Xie et al., 2021 or lack of replication of Bradlow & Bent, 2008 on foreign-accented speech). Altogether, it therefore appears that high variability during phonetic training is beneficial for learning and generalization, but that this variability does not necessarily need to originate from various speakers as long as multiple tokens of the target phonemes are provided. In this regard, studies showed that increasing the acoustic variability of temporal or spectral cues that are irrelevant to non-native speech sounds, or adding noise (e.g. speech-shaped noise, multi-talker babble) to the training stimuli can also boost learning (Cooke & Garcia Lecumberri, 2018; P. Iverson et al., 2005; Leong et al., 2018; Ylinen et al., 2010; Y. Zhang et al., 2009). Chinese native adults for instance learnt the English vowel /i/-/ɪ/ contrast better in a modified HPVT design, where acoustic stimuli were temporally exaggerated compared to a canonical HVPT paradigm, despite this temporal manipulation was not informative to distinguish the vowel categories (Cheng et al., 2019; see also Zhang, Cheng, Qin, et al., 2021 for

a follow-up study). The authors suggested that adding the irrelevant durational cue during training reallocated learners' attention to the relevant spectral categorical information, which was better extracted, thus improving learning. Phonetic training in noise also proved to benefit foreign phoneme identification. In the HVPT study by Mi and colleagues (2021), Chinese native speakers who learnt English vowels embedded in a multi-talker babble or presented in quiet (i.e. without noise) outperformed a control group who did not benefit from any training. However, only the group trained with the babble maintained their level of performance three months after training. The benefit of the babble training was even stronger for vowel identification in different types of noise, namely temporally-modulated and babble noise. By contrast, the group trained without noise outperformed the control group only in the temporally-modulated noise condition and this effect vanished three months later. Hence, adding background noise during training can help developing more robust speech representations in the non-native language. According to Mi and colleagues (2021), this may be explained by enhanced top-down attentional processes and/or increased weight of important acoustic cues (in line with Cheng et al., 2019 and Zhang et al., 2021). Given the functional role of motor regions in challenging speech perception (see Section II), it is also possible, although this was not discussed by the authors, that training in noise may favor the reliance on motor forward internal models that would benefit non-native phoneme categorization. Additional work is needed to further assess this issue, both on foreign speech sound perception and production (see Mora et al., 2022 for a study on production with HVPT in noise).

The above-reviewed HVPT studies focused on purely auditory training, leaving aside visual articulatory information available from lip-reading, that otherwise plays an important role in face to face communication (Dohen et al., 2010; Hardison & Pennington, 2021; Mcgurk & Macdonald, 1976). A few other studies have compared the effectiveness of audiovisual and auditory training, and most of them showed an advantage of providing additional visual cues on the perception and/or production of newly learnt foreign phonemes (Hardison, 2003, 2005; Hazan et al., 2005; Inceoglu, 2016; Y. Li & Somlak, 2019; X. Wang et al., 2014). Pereira Reyes and Hazan (2021) yet found comparable improvement in English vowel identification and production by Spanish native speakers following audiovisual and auditory phonetic training. Remarkably, training only on visual cues (without any auditory input) had the same effects, suggesting that merely attending to lip articulatory gestures during training can promote the learning of non-native phonemes. Other work on the other hand revealed that the efficiency of audiovisual training may depend on factors such as the informational value of the visual cues and the phonemic contrasts to acquire (Hazan et al., 2006; Ortega-Llebaria et al., 2001; Werker et al., 1992). In their HVPT study, Hazan and colleagues (2005) showed that audiovisual training in Japanese learners benefitted phonemic identification more than auditory training for the labial/labiodental /p/-/v/ contrast for which visual information is highly distinctive. This was not the case for the /ɹ/-/l/ alveolar contrast which is less visually salient (but see Hardison, 2003), undergoing a similar perceptual improvement after audiovisual versus auditory training. Better pronunciation of this latter contrast

was nevertheless observed after audiovisual than after auditory training, suggesting that information on articulatory gestures was sufficient to improve the production (Hazan et al., 2005). In this regard, Massaro and Light (2003) did not report any further improvement in Japanese learners of English when trained with a computer-animated talking head illustrating the internal oral cavity and the precise articulatory gestures for the /ɹ/-/l/ contrast, compared to training with a classical frontal view of the (tutor's) talking head (see Grauwinkel et al., 2007; Wik & Engwall, 2008 for supporting evidence). Hence, although multisensory training may foster non-native phonological learning, this is not always the case especially when visual articulatory information is not salient enough, either from the facial movements or depending on the learners' language experience. Considering the potential advantage of supplementary visual information and fully exploiting the embodied nature of speech, new training paradigms integrating manual gestures have emerged to overcome the lack of accessibility of relevant articulatory cues for learning non-native phonemes.

V. Embodied training for learning foreign phonemes

Spontaneous hand gestures come along with speech in all languages and cultures, providing complementary meaning to the auditory verbal input (Goldin-Meadow & Alibali, 2013; Iverson & Goldin-Meadow, 1998; Iverson & Thelen, 1999; McNeill, 1992, 2000; Wagner et al., 2014). This intertwining between speech and gestures arises early during native language development (Goldin-Meadow, 2010; Iverson, 2010) and gestures keep on easing language production later on in healthy adults and in patients with language and communication disorders (Akbıyık et al., 2018; Clough & Duff, 2020; Hogrefe et al., 2013). Gestures have also been shown to benefit vocabulary learning in a foreign language, mostly when they illustrate the semantic content of target words (i.e. iconic gestures; Gullberg, 2006; Kelly & Lee, 2012; Macedonia, 2014; Macedonia & Klimesch, 2014; and Kühne & Gianelli, 2019 for a review). The last decade has seen a growing interest in gestural learning for foreign phonemes, however mixed results have been reported (e.g. Amand & Touhami, 2016; Baills et al., 2019; Hirata Yukari et al., 2014; Li et al., 2020, 2021; Xi et al., 2020; Zheng et al., 2018).

Several studies demonstrated that manual pitch gestures, mimicking the fundamental frequency (F0) contour of speech, can facilitate word learning in non-native tonal languages. Learners who observed and/or imitated upward and downward hand gestures to depict, respectively, high- and low-frequency pitch contours during training indeed improved their perception or pronunciation of lexical tones (Baills et al., 2019; Zhen et al., 2019; Zheng et al., 2018; see also Hannah et al., 2016, 2017 for perception paradigms without any training). Morett and Chang (2015) yet failed to show any gain in Mandarin tone identification in English native speakers trained by imitating pitch gestures compared to a non-gestural training. A subsequent word-meaning association task however showed better performance in the gestural condition, supporting the advantage of metaphorical pitch gestures in learning foreign words that differ in lexical tones (see also Morett, 2023 for an

EEG study). It is also of note that enacting pitch gestures might not be more beneficial to tone learning than merely observing them, as shown by the few studies that directly compared the two modalities (Baills et al., 2019). Still, at the suprasegmental level, the beneficial effects of arm/hand gestures were shown on the perception (Kelly et al., 2017) and pronunciation (Yuan et al., 2019) of intonational patterns, as well as on the accentedness of foreign speech (Baills et al., 2018; Baills & Prieto, 2023; Gluhareva & Prieto, 2016). In Kushch's work (2018), Catalan learners produced Russian words with a better accent, as evaluated by Russian native speakers, after training that involved beat gestures highlighting speech prominence. This was particularly the case if the gestures had been imitated rather than observed. Along the same line, Baills and colleagues (2022) found that Catalan learners improved in French accent in an oral reading task after training with sentence-level prosodic (pitch) gestures (but see Baills, Santiago, et al., 2022 for contradictory findings).

Besides prosodic patterns, embodied training paradigms have also been developed to encode segmental information such as vowel-length contrasts. Within this scope, beat and durational gestures have mostly been used to train discrimination between short and long vowels, respectively, but consensual evidence for their benefits is so far lacking. Whereas beat gestures (McNeill, 1992) consist in non-referential up-and-down movements associated with prosodic prominence, durational gestures are typically represented with horizontal hand-sweep movements. Hirata and Kelly (2010) investigated the effect of lip movements and/or hand gestures in English native adults learning Japanese vowel-length contrasts such as /i/-/iː/. Four types of trainings (4 sessions over 2 weeks with immediate feedback) were proposed: (1) auditory input, (2) auditory input and visual lip movements, (3) auditory input and visual hand gestures, or (4) auditory input and both visual lip and hand gestures. In the two hand gestural conditions, the instructor produced short and long vowels concurrently with, respectively, a hand flick (beat gesture) and a prolonged horizontal hand sweep (durational gesture) that the participants had to observe. Results revealed better vowel identification in all training groups, but with larger improvement after the audio-lip training. Hence, providing hand gestures during training did not particularly help learners in perceiving the length difference between vowels (see also Hirata et al., 2014; Kelly et al., 2017 and Kelly & Hirata, 2017 for similar conclusions). One possible interpretation for these findings is that mixing the two types of gestures during training may have prompted learners to focus more on gesture discrimination than on the auditory speech input. The lack of obvious correspondence between the hand flick gesture and the short vowel, as well as the possibility that beat gestures may benefit suprasegmental processing in the native language but not non-native segmental processing (Hubbard et al., 2009; Krahmer & Swerts, 2007), may also account for the poor efficiency of the hand gestures in this study. Highlighting the importance of targeting the right gestures, Li and colleagues (2020) revealed that imitating horizontal hand-sweep gestures whose duration mimicked vowel length during training improved the distinction of Japanese short and long vowels (/e/-/eː/ and /o/-/oː/) by Catalan adults. This advantage of durational gestures over training without gestures was especially found on the production

of the non-native phonemes, whereas the two types of training led to similar enhancement of identification performance (see also Li et al., 2023 for effects of prosodic gestures on the production of French front-rounded vowels by Catalan speakers). Despite these encouraging results and the fact that durational gestures are spontaneously used to teach foreign language pronunciation in classrooms (Smotrova, 2017 for a review), further empirical evidence is therefore needed to fully support the beneficial role of hand gestures on the learning of non-native durational vowel contrasts.

What about phonetic features? Can manual gestures that explicitly code for place or manner of articulation help learning non-native speech sounds? A handful of recent studies have tackled this issue, with generally promising results, at least on the production side (e.g. Amand & Touhami, 2016 for unreleased plosives; Ozakin et al., 2023 for fricatives; Xi et al., 2023 for vowel lip aperture). Xi and colleagues (2020) trained Catalan adults to learn Chinese plosive and affricate consonants contrasting on aspiration, either while observing a fist-to-open hand gesture illustrating the extra air burst for aspiration, or without any manual gestures. Notably, the fist-to-open hand gesture closely mimicked the production (and perception) of the aspirated plosives (sudden opening of the fingers illustrating the quick opening of the lips and prominent air burst), whereas it less well matched the aspirated affricates characterized by a more gradual and less prominent air release. After a five-minute training session without any feedback, results revealed better pronunciation of the aspirated plosives only in the gesture group. No gestural advantage was found for the aspirated affricates. Along with previous work, identification performance for both plosives and affricates did not benefit from hand gestures compared to the no-gesture training condition. These findings emphasize that only manual gestures that appropriately reflect the phonetic features of non-native phonemes can foster their acquisition in adults (in line with Hirata & Kelly, 2010 for beat gestures mismatching short vowels). A follow-up study (Li et al., 2021) confirmed the importance not only of adding gestures during training but also of the accuracy of the learners' gestural performance. Catalan adults who appropriately imitated bimanual fist-to-open hand gestures while repeating Mandarin aspirated plosives during training indeed improved better on uttering these phonemes than learners who poorly imitated the same hand gestures. This was reflected by enhanced voice onset time (VOT) values and better rating of the trainees' pronunciation by Mandarin native speakers in the well-performed gesture group at post-test (immediately following the one-session training) as well as three days later. By contrast, in the poorly-performed gesture group, VOT did not change at post-test and the benefit of hand gestures on the rated pronunciation was no longer seen at the delayed post-test. The quality of the imitated gestures is therefore crucial to yield positive effects of embodied training on learning and maintaining non-native phoneme production, pointing to the need of assessing learners' gestural performance as well as of designing paradigms with adequate gestures.

The complexity of the manual gestures and the fact that they stand for visible or non-visible articulatory features also seems to impact learning efficiency. In the study by Hoetjes and van Maastricht (2020), Spanish

adults learnt to produced two Dutch phonemes that are part (/u/) or not (/θ/) of their native repertoire and that require new phoneme-grapheme correspondences to be-acquired. The easy vowel /u/ was better produced after training based on the observation of an iconic hand gesture illustrating the rounding of the lips rather than on the observation of a simple pointing gesture to the mouth. The reverse was found for the more challenging consonant /θ/: learning was hindered by an iconic gesture indicating to push the tongue between the teeth while it was more efficient with the pointing gesture. The authors suggested that manual gestures reflecting phonetic features may help phonemic learning only when processing demands are not too high, such as for the easy vowel /u/. When processing cost increases, to acquire a non-native phoneme outside the native phonological inventory for instance, providing complex hand gestures may be detrimental to learning (see Kelly & Lee, 2012 for similar arguments). It is of note that even though Hoetjes and van Maastricht (2020) did not discuss this point, the fact that the gestures illustrated the lips or the tongue, namely articulators that are directly visible or not for the learners, may also have affected the effectiveness of learning. As a matter of fact, Xi and coworkers (2024) revealed that observing lip-related gestures during training facilitated the production of the English vowels /æ/ and /ʌ/ by Catalan-Spanish adults more than observing gestures mimicking tongue shape within the mouth. These two vowels differ in both the degree of lip aperture and tongue position along the antero-posterior plane, and they tend to be assimilated to /a/ by Spanish speakers. Gestural training therefore involved either a one-handed gesture depicting the lip aperture needed to produce the vowels, or a bimanual gesture representing tongue backness relative to a reference point (as well as lip aperture from the distance between the two hands). A control group was trained in a classical audiovisual condition without hand gestures. Identification improved comparably in all training groups, in line with the limited effects of gestural paradigms on perception (Li et al., 2021; Xi et al., 2020). For production however, results revealed that the lip-related gestures helped more the learners to adjust their lip aperture (as measured by formant values) for non-native vowels than the tongue-related gesture and non-gestural conditions. The efficacy of the training to adjust tongue position was on the contrary limited and similar across the three groups. Hence, hand gestures that encode visible articulatory features, such as lip aperture, may be more beneficial than gestures coding for non-visible features, involving the tongue in particular, as the latter may potentially increase the processing demands. Indeed, in line with previous work, manual gestures mimicking the tongue shape do not match visual facial information and may therefore create some kind of incongruency for the learners as opposed to lip-related gestures that give complementary congruent information about the way phonemes are produced. Notably, as pointed out by Xi and colleagues (2024), the lack of feedback in the training paradigm in their study, as well as in the work by Hoetjes and van Maastricht (2020), may explain the limited learning advantage of tongue-related hand gestures. As a matter of fact, two studies, in a classroom (Lan & Wu, 2013) and in a clinical setting (Rusiewicz & Rivera, 2017), reported better non-native or native consonant pronunciation after teaching with hand gestures that illustrated the shape of

the tongue. The fact that learners only observed the manual gestures in the two former studies but actually imitated them in the latter two may also be a key ingredient for efficient learning (in line with studies on vocabulary learning or more general cognitive skills; e.g. Goldin-Meadow et al., 2009; Macedonia et al., 2011). While non-visible articulatory gestures may be challenging to process due to incongruency with visual facial cues, they might still be effective when paired with active imitation, highlighting the importance of embodied practices for enhancing phonetic acquisition.

Overall, current training paradigms offer a limited framework that varies in effectiveness with regards to various gestures and the few phonemes investigated across different foreign languages. Yet, there are some implications for future training paradigms to build new phonemic categories so as to improve both perception and production. Gestures that emphasize perceptual distinctions between phonemes should be concise and accurately represent the articulatory features of the target foreign speech sounds. When articulatory features are not directly visible (e.g., tongue position or articulatory gesture within the oral cavity), a training paradigm embedding gesture imitation seems recommended to enhance learning. Assessing the motor performance of these gestures during the training phase could also be crucial to maximize learning efficacy. In addition, longer training paradigms, currently absent in the literature, may be beneficial in strengthening the link between gestures and perceived articulatory features, thereby further improving the perception and production of the foreign phonemes.

Conclusion

We here provided an overview of the literature on how the motor system contributes to both native and non-native speech perception, as well as how learning non-native speech sounds can benefit from multisensory support. Current evidence indicates that the (pre)motor regions involved in speech production are also activated during speech perception, underscoring a foundational mechanism of sensorimotor processing in speech. Specific motor cortical activity linked to distinct articulatory features further supports the embodied nature of phoneme perception. This motor resonance occurs particularly under challenging perceptual conditions when auditory information is degraded, as well as in the context of non-native speech. Given the motor system's contribution to decode subtle articulatory features in perceived non-native phonemes, the potential benefits of sensorimotor-based training for learning become evident. Training paradigms that incorporate high variability in phoneme tokens, especially acoustically varied with noise, can enhance perceptual skills across different phonetic contrasts in foreign languages. Multisensory protocols for learning, such as training with manual gestures, further support the acquisition of non-native speech sounds and have been found to especially improve the production of foreign phonemes. However, the limited number of studies that have successfully trained phonetic features to learn foreign phonemes may restrict the observable

benefits of gestural learning in perception. Future research should focus on longitudinal studies to assess whether both the perception and production of foreign speech benefit from gestural training. In addition, the lack of neuroimaging studies on this topic leaves a critical gap in understanding how gestural training may fine-tune articulatory representations within the motor cortex, specifically in relation to non-native phoneme processing. This, in turn, allows for the refinement of training protocols to optimize the learning process. Studies examining developmental and cross-linguistic variations could also provide valuable insights into the effectiveness of gestural training for foreign language acquisition.

Acknowledgements

This project was supported by the French National Research Agency (ANR) for the AnchorFL project (ANR-19-CE28-0015; PIs: V. Boulenger, A. C. Roy and C. Brozzoli) and by the Appel à Projets Pluridisciplinaires Interne (APPI) from Université lyon 2 (to V. Boulenger, A. C. Roy and C. Brozzoli).

References

Adank, P., & Devlin, J. T. (2010). On-line plasticity in spoken sentence comprehension: Adapting to time-compressed speech. *NeuroImage*, *49*(1), 1124–1132. https://doi.org/10.1016/j.neuroimage.2009.07.032

Akbıyık, S., Karaduman, A., Göksun, T., & Chatterjee, A. (2018). The relationship between co-speech gesture production and macrolinguistic discourse abilities in people with focal brain injury. *Neuropsychologia*, *117*, 440–453. https://doi.org/10.1016/j.neuropsychologia.2018.06.025

Alho, J., Lin, F.-H., Sato, M., Tiitinen, H., Sams, M., & Jääskeläinen, I. P. (2014). Enhanced neural synchrony between left auditory and premotor cortex is associated with successful phonetic categorization. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00394

Amand, M., & Touhami, Z. (2016). Teaching the pronunciation of sentence final and word boundary stops to French learners of English: Distracted imitation versus audio-visual explanations. *Research in Language*, *14*(4), 4. https://doi.org/10.1515/rela-2016-0020

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P. W., Kranen-Mastenbroek, V. H. van, & Jansma, B. M. (2018). Sensorimotor Representation of Speech Perception. Cross-Decoding of Place of Articulation Features during Selective Attention to Syllables in 7T fMRI. *ENeuro*, *5*(2). https://doi.org/10.1523/ENEURO.0252-17.2018

Arsenault, J. S., & Buchsbaum, B. R. (2016). No evidence of somatotopic place of articulation feature mapping in motor cortex during passive speech perception. *Psychonomic Bulletin & Review*, *23*(4), 1231–1240. https://doi.org/10.3758/s13423-015-0988-z

Baills, F., Alazard-Guiu, C., & Prieto, P. (2022). Embodied Prosodic Training Helps Improve Accentedness and Suprasegmental Accuracy. *Applied Linguistics*, *43*(4), 776–804. https://doi.org/10.1093/applin/amac010

Baills, F., & Prieto, P. (2023). Embodying rhythmic properties of a foreign language through hand-clapping helps children to better pronounce words. *Language Teaching Research*, *27*(6), 1576–1606. https://doi.org/10.1177/1362168820986716

Baills, F., Santiago, F., Mairano, P., & Prieto, P. (2022). The effects of prosodic training with logatomes and prosodic gestures on L2 spontaneous speech. In *Speech Prosody 2022*. ISCA. https://doi.org/10.21437/SpeechProsody.2022-163

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). OBSERVING AND PRODUCING PITCH GESTURES FACILITATES THE LEARNING OF MANDARIN CHINESE TONES AND WORDS. *Studies in Second Language Acquisition*, *41*(1), 33–58. https://doi.org/10.1017/S0272263118000074

Baills, F., Zhang, Y., & Prieto, P. (2018). *Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from Catalan and Chinese learners of French*. 853–857. https://doi.org/10.21437/SpeechProsody.2018-172

Bartoli, E., D'Ausilio, A., Berry, J., Badino, L., Bever, T., & Fadiga, L. (2015). Listener–Speaker Perceived Distance Predicts the Degree of Motor Contribution to Speech Perception. *Cerebral Cortex*, *25*(2), 281–288. https://doi.org/10.1093/cercor/bht257

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

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*(4), 2299–2310.

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

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

Callan, D., Callan, A., Gamez, M., Sato, M., & Kawato, M. (2010). Premotor cortex mediates perceptual performance. *NeuroImage*, *51*(2), 844–858. https://doi.org/10.1016/j.neuroimage.2010.02.027

Callan, D., Callan, A., & Jones, J. A. (2014). Speech motor brain regions are differentially recruited during perception of native and foreign-accented phonemes for first and second language listeners. *Frontiers in Neuroscience*, *8*, 275. https://doi.org/10.3389/fnins.2014.00275

Callan, D. E., Jones, J. A., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, *22*(3), 1182–1194. https://doi.org/10.1016/j.neuroimage.2004.03.006

Callan, D. E., Tajima, K., Callan, A. M., Kubo, R., Masaki, S., & Akahane-Yamada, R. (2003). Learning-induced neural plasticity associated with improved identification performance after training of a difficult second-language phonetic contrast. *NeuroImage*, *19*(1), 113–124. https://doi.org/10.1016/S1053-8119(03)00020-X

Callan, D., Jones, J. A., Munhall, K., Callan, A. M., Kroos, C., & Vatikiotis-Bateson, E. (2003). Neural processes underlying perceptual enhancement by visual speech gestures. *NeuroReport*, *14*(17), 2213.

Callan, D., Jones, J., Callan, A. M., & Akahane-Yamada, R. (2004). Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory–auditory/orosensory internal models. *NeuroImage*, *22*(3), 1182–1194. https://doi.org/10.1016/j.neuroimage.2004.03.006

Cebrian, J., & Carlet, A. (2014). Second-Language Learners' Identification of Target-Language Phonemes: A Short-Term Phonetic Training Study. *The Canadian Modern Language Review*, *70*(4), 474–499. https://doi.org/10.3138/cmlr.2318

Cheng, B., Zhang, X., Fan, S., & Zhang, Y. (2019). The Role of Temporal Acoustic Exaggeration in High Variability Phonetic Training: A Behavioral and ERP Study. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01178

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *Elife*, *5*, e12577.

Choi, D., Bruderer, A. G., & Werker, J. F. (2019). Sensorimotor influences on speech perception in pre-babbling infants: Replication and extension of Bruderer et al. (2015). *Psychonomic Bulletin & Review*, *26*(4), 1388–1399. https://doi.org/10.3758/s13423-019-01601-0

Clough, S., & Duff, M. C. (2020). The Role of Gesture in Communication and Cognition: Implications for Understanding and Treating Neurogenic Communication Disorders. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.00323

Cooke, M., & Garcia Lecumberri, M. L. (2018). Effects of exposure to noise during perceptual training of non-native language sounds. *The Journal of the Acoustical Society of America*, *143*(5), 2602–2610. https://doi.org/10.1121/1.5035080

Correia, J. M., Jansma, B. M. B., & Bonte, M. (2015). Decoding Articulatory Features from fMRI Responses in Dorsal Speech Regions. *Journal of Neuroscience*, *35*(45), 15015–15025. https://doi.org/10.1523/JNEUROSCI.0977-15.2015

D'Ausilio, A., Bufalari, I., Salmas, P., & Fadiga, L. (2012). The role of the motor system in discriminating normal and degraded speech sounds. *Cortex*, *48*(7), 882–887. https://doi.org/10.1016/j.cortex.2011.05.017

D'Ausilio, A., Maffongelli, L., Bartoli, E., Campanella, M., Ferrari, E., Berry, J., & Fadiga, L. (2014). Listening to speech recruits specific tongue motor synergies as revealed by transcranial magnetic stimulation and tissue-Doppler ultrasound imaging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *Volume 369*(Issue 1644).

D'Ausilio, A., Pulvermüller, F., Salmas, P., Bufalari, I., Begliomini, C., & Fadiga, L. (2009). The Motor Somatotopy of Speech Perception. *Current Biology*, *19*(5), 381–385. https://doi.org/10.1016/j.cub.2009.01.017

Dohen, M., Schwartz, J.-L., & Bailly, G. (2010). Speech and face-to-face communication – An introduction. *Speech Communication*, *52*(6), 477–480. https://doi.org/10.1016/j.specom.2010.02.016

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences*, *111*(19), 7126–7131. https://doi.org/10.1073/pnas.1318738111

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2016). Increased activity in frontal motor cortex compensates impaired speech perception in older adults. *Nature Communications*, *7*(1), 12241. https://doi.org/10.1038/ncomms12241

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*(2), 399–402. https://doi.org/10.1046/j.0953-816x.2001.01874.x

Fischer, M. H., & Zwaan, R. A. (2008). Embodied Language: A Review of the Role of the Motor System in Language Comprehension. *Quarterly Journal of Experimental Psychology*, *61*(6), 825–850. https://doi.org/10.1080/17470210701623605

Franken, M. K., Liu, B. C., & Ostry, D. J. (2022). Towards a somatosensory theory of speech perception. *Journal of Neurophysiology*, *128*(6), 1683–1695. https://doi.org/10.1152/jn.00381.2022

Gluhareva, D., & Prieto, P. (2016). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, *21*. https://doi.org/10.1177/1362168816651463

Goldin-Meadow, S. (2010). GESTURE'S ROLE IN CREATING AND LEARNING LANGUAGE. *Enfance; Psychologie, Pedagogie, Neuropsychiatrie, Sociologie*, *2010*(3), 239. https://doi.org/10.4074/S0013754510003034

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's Role in Speaking, Learning, and Creating Language. *Annual Review of Psychology*, *64*(Volume 64, 2013), 257–283. https://doi.org/10.1146/annurev-psych-113011-143802

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing Gives Children New Ideas About Math. *Psychological Science*, *20*(3), 267. https://doi.org/10.1111/j.1467-9280.2009.02297.x

Golestani, N., & Zatorre, R. J. (2004). Learning new sounds of speech: Reallocation of neural substrates. *NeuroImage*, *21*(2), 494–506. https://doi.org/10.1016/j.neuroimage.2003.09.071

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2), 55–67. https://doi.org/10.1016/j.bandl.2008.01.005

Grauwinkel, K., Dewitt, B., & Fagel, S. (2007). *Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech*. 706–709. https://doi.org/10.21437/Interspeech.2007-295

Gullberg, M. (2006). *Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon)*. *44*(2), 103–124. https://doi.org/10.1515/IRAL.2006.004

Hannah, B., Wang, Y., Jongman, A., & Sereno, J. A. (2016). Cross-modal association between auditory and visual-spatial information in Mandarin tone perception. *The Journal of the Acoustical Society of America*, *140*(4_Supplement), 3225. https://doi.org/10.1121/1.4970187

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal Association between Auditory and Visuospatial Information in Mandarin Tone Perception in Noise by Native and Non-native Perceivers. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02051

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522. https://doi.org/10.1017/S0142716403000250

Hardison, D. M. (2005). Second-language spoken word identification: Effects of perceptual training, visual cues, and phonetic environment. *Applied Psycholinguistics*, *26*(4), 579–596. https://doi.org/10.1017/S0142716405050319

Hardison, D. M., & Pennington, M. C. (2021). Multimodal Second-Language Communication: Research Findings and Pedagogical Implications. *RELC Journal*, *52*(1), 62–76. https://doi.org/10.1177/0033688220966635

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., & Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrastsa). *The Journal of the Acoustical Society of America*, *119*(3), 1740–1751. https://doi.org/10.1121/1.2166611

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

Hervais-Adelman, Carlyon, Johnsrude, & Davis. (2012). *Brain regions recruited for the effortful comprehension of noise-vocoded words*. https://www.tandfonline.com/doi/epdf/10.1080/01690965.2012.662280?needAccess=true

Hickok, G. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hincapié Casas, A. S., Lajnef, T., Pascarella, A., Guiraud-Vinatea, H., Laaksonen, H., Bayle, D., Jerbi, K., & Boulenger, V. (2021). Neural oscillations track natural but not artificial fast speech: Novel insights from speech-brain coupling using MEG. *NeuroImage*, *244*, 118577. https://doi.org/10.1016/j.neuroimage.2021.118577

Hirata, Y., & Kelly, S. D. (2010). Effects of Lips and Hands on Auditory Learning of Second-Language Speech Sounds. *Journal of Speech, Language, and Hearing Research*, *53*(2), 298–310. https://doi.org/10.1044/1092-4388(2009/08-0243)

Hirata Yukari, Kelly Spencer D., Huang Jessica, & Manansala Michael. (2014). Effects of Hand Gestures on Auditory Learning of Second-Language Vowel Length Contrasts. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2090–2101. https://doi.org/10.1044/2014_JSLHR-S-14-0049

Hoetjes, M., & van Maastricht, L. (2020). Using Gesture to Facilitate L2 Phoneme Acquisition: The Importance of Gesture and Phoneme Complexity. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.575032

Hogrefe, K., Ziegler, W., Wiesmayer, S., Weidinger, N., & Goldenberg, G. (2013). The actual and potential use of gestures for communication in aphasia. *Aphasiology*, *27*(9), 1070–1089. https://doi.org/10.1080/02687038.2013.803515

Hubbard, A. L., Wilson, S. M., Callan, D. E., & Dapretto, M. (2009). Giving speech a hand: Gesture modulates activity in auditory cortex during speech perception. *Human Brain Mapping*, *30*(3), 1028–1037. https://doi.org/10.1002/hbm.20565

Iacoboni, M. (2008). The role of premotor cortex in speech perception: Evidence from fMRI and rTMS. *Journal of Physiology-Paris*, *102*(1–3), 31–34. https://doi.org/10.1016/j.jphysparis.2008.03.003

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, *37*(5), 1175–1199. https://doi.org/10.1017/S0142716415000533

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences*, *106*(4), 1245–1248. https://doi.org/10.1073/pnas.0810063106

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language*, *37*(2), 229–261. https://doi.org/10.1017/S0305000909990432

Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, *396*(6708), 228–228. https://doi.org/10.1038/24300

Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, *6*(11–12), 19–40.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, *122*(5), 2842–2854. https://doi.org/10.1121/1.2783198

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakersa). *The Journal of the Acoustical Society of America*, *126*(2), 866–877. https://doi.org/10.1121/1.3148196

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*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160. https://doi.org/10.1017/S0142716411000300

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric Gestures Facilitate Perception of Intonation More than Length in Auditory Judgments of Non-Native Phonemic Contrasts. *Collabra: Psychology*, *3*(1), 7. https://doi.org/10.1525/collabra.76

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807. https://doi.org/10.1080/01690965.2011.581125

Kelly, S., & Hirata, Y. (2017). *What neural measures reveal about foreign language learning of Japanese vowel length contrasts with hand gestures.* (pp. 278–294).

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Kühne, K., & Gianelli, C. (2019). Is embodied cognition bilingual? Current evidence and perspectives of the embodied cognition approach to bilingual language processing. *Frontiers in Psychology*, 10, 108.

Kushch, O. (2018). Beat gestures and prosodic prominence: Impact on learning [Ph.D. Thesis, Universitat Pompeu Fabra]. In *TDX (Tesis Doctorals en Xarxa)*. https://www.tdx.cat/handle/10803/463004

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247. https://doi.org/10.1017/S0142716405050150

Lan, Y., & Wu, M. (2013). Application of Form-Focused Instruction in English Pronunciation: Examples from Mandarin Learners. *Creative Education*, *04*(09), 09. https://doi.org/10.4236/ce.2013.49B007

Leong, C. X. R., Price, J. M., Pitchford, N. J., & Heuven, W. J. B. van. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLOS ONE*, *13*(10), e0204888. https://doi.org/10.1371/journal.pone.0204888

Li, P., Bailles, F., Baqué, L., & Prieto, P. (2023). The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second Language Research*, *39*(4), 1077–1105. https://doi.org/10.1177/02676583221124075

Li, P., Bailles, F., & Prieto, P. (2020). OBSERVING AND PRODUCING DURATIONAL HAND GESTURES FACILITATES THE PRONUNCIATION OF NOVEL VOWEL-LENGTH CONTRASTS. *Studies in Second Language Acquisition*, *42*(5), 1015–1039. https://doi.org/10.1017/S0272263120000054

Li, P., Xi, X., Baills, F., & Prieto, P. (2021). Training non-native aspirated plosives with hand gestures: Learners' gesture performance matters. *Language, Cognition and Neuroscience*, *36*(10), 1313–1328. https://doi.org/10.1080/23273798.2021.1937663

Li, Y., & Somlak, T. (2019). The effects of articulatory gestures on L2 pronunciation learning: A classroom-based study. *Language Teaching Research*, *23*(3), 352–371. https://doi.org/10.1177/1362168817730420

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*(3), 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*(4), 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*(2), 874–886. https://doi.org/10.1121/1.1894649

Lotto, A. J., Hickok, G. S., & Holt, L. L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, *13*(3), 110–114. https://doi.org/10.1016/j.tics.2008.11.008

Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01467

Macedonia, M., & Klimesch, W. (2014). Long-Term Effects of Gestures on Memory for Foreign Language Words Trained in the Classroom. *Mind, Brain, and Education*, *8*(2), 74–88. https://doi.org/10.1111/mbe.12047

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*(6), 982–998. https://doi.org/10.1002/hbm.21084

McClelland, J. L., Fiez, J. A., & McCandliss, B. D. (2002). Teaching the /r/–/l/ discrimination to Japanese adults: Behavioral and neural aspects. *Physiology & Behavior*, *77*(4), 657–662. https://doi.org/10.1016/S0031-9384(02)00916-2

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

McNeill, D. (2000). *Language and Gesture*. Cambridge University Press.

Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The Essential Role of Premotor Cortex in Speech Perception. *Current Biology*, *17*(19), 1692–1696. https://doi.org/10.1016/j.cub.2007.08.064

Mi, L., Tao, S., Wang, W., Dong, Q., Dong, B., Li, M., & Liu, C. (2021). Training non-native vowel perception: In quiet or noise. *The Journal of the Acoustical Society of America*, *149*(6), 4607–4619. https://doi.org/10.1121/10.0005276

Mora, J. C., Ortega, M., Mora-Plaza, I., & Aliaga-García, C. (2022). Training the pronunciation of L2 vowels under different conditions: The use of non-lexical materials and masking noise. *Phonetica*, *79*(1), 1–43. https://doi.org/10.1515/phon-2022-2018

Morett, L. M. (2023). Observing gesture at learning enhances subsequent phonological and semantic processing of L2 words: An N400 study. *Brain and Language*, *246*, 105327. https://doi.org/10.1016/j.bandl.2023.105327

Morett, L. M., & Chang, L.-Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Möttönen, R., & Watkins, K. E. (2009). Motor Representations of Articulators Contribute to Categorical Perception of Speech Sounds. *Journal of Neuroscience*, *29*(31), 9819–9825. https://doi.org/10.1523/JNEUROSCI.6018-08.2009

Murakami, T., Kell, C. A., Restle, J., Ugawa, Y., & Ziemann, U. (2015). Left dorsal speech stream components and their contribution to phonological processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *35*(4), 1411–1422. https://doi.org/10.1523/JNEUROSCI.0246-14.2015

Nishi, K., & Kewley, -Port Diane. (2007). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Nishi, K., & Kewley-Port, D. (2008). Non-native Speech Perception Training Using Vowel Subsets: Effects of Vowels in Sets and Order of Training. *Journal of Speech, Language, and Hearing Research : JSLHR*, *51*(6), 1480–1493. https://doi.org/10.1044/1092-4388(2008/07-0109)

Nuttall, H. E., Kennedy-Higgins, D., Devlin, J. T., & Adank, P. (2017). The role of hearing ability and speech distortion in the facilitation of articulatory motor cortex. *Neuropsychologia*, *94*, 13–22. https://doi.org/10.1016/j.neuropsychologia.2016.11.016

Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., & Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *NeuroImage*, *128*, 218–226. https://doi.org/10.1016/j.neuroimage.2015.12.038

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). *Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English* [Proceedings paper]. In: Massaro, DW and Light, J and Geraci, K, (Eds.) Auditory-Visual Speech Processing (AVSP 2001). (Pp. 149 - 154). Auditory-Visual Speech Association (2001); Auditory-Visual Speech Association. http://www.isca-speech.org/archive_open/avsp01/av01_149.html

Osnes, B., Hugdahl, K., & Specht, K. (2011). Effective connectivity analysis demonstrates involvement of premotor cortex during speech perception. *NeuroImage*, *54*(3), 2437–2445. https://doi.org/10.1016/j.neuroimage.2010.09.078

Ozakin, A. S., Xi, X., Li, P., & Prieto, P. (2023). Thanks or Tanks: Training with Tactile Cues Improves Learners' Accuracy of English Interdental Consonants in an Oral Reading Task. *Language Learning and Development*, *19*(4), 404–419. https://doi.org/10.1080/15475441.2022.2107522

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472. https://doi.org/10.1121/1.3593366

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. *The Journal of the Acoustical Society of America*, *119*(3), 1684–1696. https://doi.org/10.1121/1.2161427

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 5. https://doi.org/10.1038/nrn2811

Pulvermuller, F., Huss, M., Kherif, F., Moscoso del Prado Martin, F., Hauk, O., & Shtyrov, Y. (2006). Motor cortex maps articulatory features of speech sounds. *Proceedings of the National Academy of Sciences*, *103*(20), 7865–7870. https://doi.org/10.1073/pnas.0509989103

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Reyes, Y. P., & Hazan, V. (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. *Onomázein*, *51*, 51. https://doi.org/10.7764/onomazein.51.04

Roy, A. C., Craighero, L., Fabbri-Destro, M., & Fadiga, L. (2008). Phonological and lexical motor facilitation during speech listening: A transcranial magnetic stimulation study. *Journal of Physiology-Paris*, *102*(1–3), 101–105. https://doi.org/10.1016/j.jphysparis.2008.03.006

Rusiewicz, H. L., & Rivera, J. L. (2017). The Effect of Hand Gesture Cues Within the Treatment of /r/ for a College-Aged Adult With Persisting Childhood Apraxia of Speech. *American Journal of Speech-Language Pathology*, *26*(4), 1236–1243. https://doi.org/10.1044/2017_AJSLP-15-0172

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01318

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224. https://doi.org/10.1017/S0142716417000418

Sato, M., Tremblay, P., & Gracco, V. L. (2009). A mediating role of the premotor cortex in phoneme segmentation. *Brain and Language*, *111*(1), 1–7. https://doi.org/10.1016/j.bandl.2009.03.002

Schmitz, J., Bartoli, E., Maffongelli, L., Fadiga, L., Sebastian-Galles, N., & D'Ausilio, A. (2019). Motor cortex compensates for lack of sensory and motor experience during auditory speech perception. *Neuropsychologia*, *128*, 290–296. https://doi.org/10.1016/j.neuropsychologia.2018.01.006

Schomers, M. R., & Pulvermüller, F. (2016). Is the Sensorimotor Cortex Relevant for Speech Perception and Understanding? An Integrative Review. *Frontiers in Human Neuroscience*, *10*. https://www.frontiersin.org/articles/10.3389/fnhum.2016.00435

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354. https://doi.org/10.1016/j.jneuroling.2009.12.004

Scott, S. K., McGettigan, C., & Eisner, F. (2009). A little more conversation, a little less action—Candidate roles for the motor cortex in speech perception. *Nature Reviews Neuroscience*, *10*(4), 4. https://doi.org/10.1038/nrn2603

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

Skipper, J. I., Devlin, J. T., & Lametti, D. R. (2017). The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language*, *164*, 77–105. https://doi.org/10.1016/j.bandl.2016.10.004

Skipper, J. I., van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing Lips and Seeing Voices: How Cortical Areas Supporting Speech Production Mediate Audiovisual Speech Perception. *Cerebral Cortex*, *17*(10), 2387–2399. https://doi.org/10.1093/cercor/bhl147

Smotrova, T. (2017). Making Pronunciation Visible: Gesture In Teaching Pronunciation. *TESOL Quarterly*, *51*(1), 59–89. https://doi.org/10.1002/tesq.276

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology. General*, *141*(2), 363–381. https://doi.org/10.1037/a0025014

Wagner, P., Malisz, Z., & Kopp, S. (2014). Gesture and speech in interaction: An overview. *Speech Communication*, *57*, 209–232. https://doi.org/10.1016/j.specom.2013.09.008

Wang, X., Hueber, T., & Badin, P. (2014, May 5). *On the use of an articulatory talking head for second language pronunciation training: The case of Chinese learners of French*.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*(2), 1033–1043. https://doi.org/10.1121/1.1531176

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

Werker, J. F., Frost, P. E., & McGuirk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *46*(4), 551–568. https://doi.org/10.1037/h0084331

Wik, P., & Engwall, O. (2008). Looking at tongues—Can it help in speech perception? *Proceedings FONETIK 2008*. https://www.academia.edu/28071988/Looking_at_tongues_can_it_help_in_speech_perception

Wilson, S. M., & Iacoboni, M. (2006). Neural responses to non-native phonemes varying in producibility: Evidence for the sensorimotor nature of speech perception. *NeuroImage*, *33*(1), 316–325. https://doi.org/10.1016/j.neuroimage.2006.05.032

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004a). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702. https://doi.org/10.1038/nn1263

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004b). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702. https://doi.org/10.1038/nn1263

Xi, X., Li, P., Baills, F., & Prieto, P. (2020). Hand Gestures Facilitate the Acquisition of Novel Phonemic Contrasts When They Appropriately Mimic Target Phonetic Features. *Journal of Speech, Language, and Hearing Research*, *63*(11), 3571–3585. https://doi.org/10.1044/2020_JSLHR-20-00084

Xi, X., Li, P., & Prieto, P. (2023). Reducing acoustic overlap of L2 English vowels through gestures encoding lip aperture. In A. Henderson & A. Kirkova-Naskova (Eds.), *Proceedings of the 7th International Conference on English Pronunciation: Issues and Practices* (pp. 261–270). https://doi.org/10.5281/zenodo.8225191

Xi, X., Li, P., & Prieto, P. (2024). Improving Second Language Vowel Production With Hand Gestures Encoding Visible Articulation: Evidence From Picture-Naming and Paragraph-Reading Tasks. *Language Learning*, *n/a*(n/a), 1–33. https://doi.org/10.1111/lang.12647

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General*, *150*(11), e22–e56. https://doi.org/10.1037/xge0001039

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*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Yuan, C., González-Fuente, S., Baills, F., & Prieto, P. (2019). OBSERVING PITCH GESTURES FAVORS THE LEARNING OF SPANISH INTONATION BY MANDARIN SPEAKERS. *Studies in Second Language Acquisition*, *41*(1), 5–32. https://doi.org/10.1017/S0272263117000316

Yuen, I., Davis, M. H., Brysbaert, M., & Rastle, K. (2010). Activation of articulatory information in speech perception. *Proceedings of the National Academy of Sciences*, *107*(2), 592–597. https://doi.org/10.1073/pnas.0904774107

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, *87*, 101071. https://doi.org/10.1016/j.wocn.2021.101071

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*(12), 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181

Zhang, Y., Kuhl, P. K., Imada, T., Iverson, P., Pruitt, J., Stevens, E. B., Kawakatsu, M., Tohkura, Y., & Nemoto, I. (2009). Neural signatures of phonetic learning in adulthood: A magnetoencephalography study. *NeuroImage*, *46*(1), 226–240. https://doi.org/10.1016/j.neuroimage.2009.01.028

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, *187*, 178–187. https://doi.org/10.1016/j.cognition.2019.03.004

Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the Effects of Imitating Hand Gestures and Head Nods on L1 and L2 Mandarin Tone Production. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2179–2195. https://doi.org/10.1044/2018_JSLHR-S-17-0481