Embodied speech: Sensorimotor contributions to native and non-native phoneme processing and learning

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

Neurobiology of Language

an open access journal

Citation: Tseng, T., Krzonowski, J., Brozzoli, C., Roy, A. C., & Boulenger, V. (2026). Embodied speech: Sensorimotor contributions to native and non-native phoneme processing and learning. *Neurobiology of Language*, *7*, NOL.a.215. https://doi.org/10.1162/NOL.a.215

DOI:
https://doi.org/10.1162/NOL.a.215

Received: 28 March 2025
Accepted: 8 October 2025

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Authors:
Tzuyi Tseng
tzuyitseng.neuroling@gmail.com
Véronique Boulenger
veronique.boulenger@cnrs.fr

Handling Editor:
Kate Watkins

The MIT Press

REVIEW ARTICLE

Embodied Speech: Sensorimotor Contributions to Native and Non-Native Phoneme Processing and Learning

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

[1]CNRS, Université Lyon 2, Laboratoire Dynamique du Langage, Lyon, France
[2]Integrative Multisensory Perception Action & Cognition Team (ImpAct), Centre de Recherche en Neurosciences de Lyon, INSERM, Université Claude Bernard 1, Lyon, France

Keywords: embodiment, foreign language learning, manual gestures, motor system, non-native phonemes, speech perception

ABSTRACT

Learning to recognize and produce foreign speech sounds can be challenging, particularly when only subtle differences distinguish these new sounds from phonemes in the native language. Functional neuroimaging evidence shows that the motor cortex is involved in speech production and in perceptual phonemic processing. This highlights the embodied nature of speech perception, predicting the potential benefits of sensorimotor-based training approaches to enhance the acquisition of foreign speech sounds. Hence, here we first review current findings on the motor contribution to not only native but also non-native phoneme perception. Available evidence has established that motor cortical activity especially shows up under non-optimal perceptual conditions, such as when native phonemes are degraded by noise or when listeners perceive non-native speech sounds. Drawing upon this evidence, we then review training paradigms that have been developed for learning foreign phonemes, with a special emphasis on those embedding manual gestures as cues to represent phonetic features of the to-be-learned speech sounds. By pointing to both strengths and caveats of available studies, this review allows to delineate a clear framework and opens perspectives to optimize foreign phoneme learning, and ultimately support perception and production.

INTRODUCTION

Over the last 15 years, research on language production and comprehension has undergone a conceptual revolution, highlighting how language may rely not only on specific brain areas but also on embodied processes underpinned by sensorimotor brain regions. A large amount of studies has underlined the sensorimotor grounding of various processes at play in language comprehension, particularly in processing action verbs and understanding sentences (see Fischer & Zwaan, 2008, and Franken et al., 2022, for reviews). However, one crucial aspect of language processing—phoneme perception—has received comparably less attention in the context of embodied cognition. Yet, motor involvement seems to play a significant role in decoding sounds within a specific linguistic code. The focus of the current review is to provide an updated overview of research findings in this domain, focusing on both native and foreign (i.e., non-native) phoneme perception. Additionally, it examines whether and how

sensorimotor-based training may enhance phonological processing and learning of foreign language speech sounds.

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 (see Figure 1 for an overview of brain imaging studies reporting (pre)motor activity during phoneme perception). Brain activity in the motor system echoes the motor theory of speech perception (Liberman et al., 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000), suggesting that speech is perceived by decoding the invariant articulatory properties of speech sounds, namely the vocal tract gestures of the speaker used to produce those sounds. According to this view, motor regions responsible for generating speech movements are also engaged during speech perception. In their seminal work testing this theory, Fadiga and colleagues (2002) applied TMS to the left motor cortex of Italian adults while they listened to words and pseudowords 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 involves only minimal tongue movements. Corroborating Liberman's view, items embedded with /rr/ elicited higher tongue motor evoked potentials (MEPs) than the other conditions, showing automatic and somatotopic activity in the motor cortex for passive speech perception, as if articulatory movements were decoded (see also Watkins et al., 2003, for lip-MEPs during listening to or viewing speech). Roy and colleagues (2008) not only replicated this

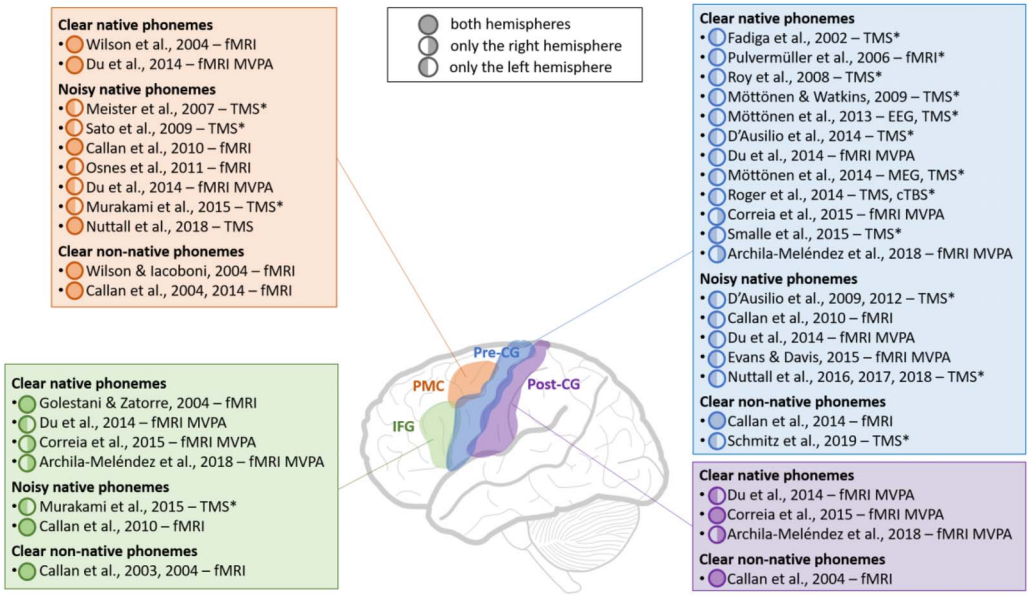

Figure 1. Overview of neuroimaging studies showing the involvement of speech production brain regions in phoneme perception. These regions include the inferior frontal gyrus (IFG, in green), premotor cortex (PMC, in orange), precentral gyrus (pre-CG, in blue), and postcentral gyrus (post-CG, in purple). For each region, the fMRI and TMS studies that investigated phoneme perception under different perceptual conditions (clear native phonemes, noisy native phonemes, and clear non-native phonemes) are indicated in a box of the corresponding color. For each study, a full-colored circle indicates that significant activity was observed in both hemispheres, whereas a half-colored circle indicates significant activity in only one hemisphere. Studies that investigated cortical activity in the left hemisphere only are marked with *. fMRI = functional magnetic resonance imaging; MVPA = multivoxel pattern analysis; TMS = transcranial magnetic stimulation; EEG = electroencephalography; MEG = magnetoencephalography; cTBS = continuous theta-burst stimulation.

phonological effect with Italian pseudowords including the tongue-related /ll/, but also revealed an early lexical effect, further supporting the link between speech perception and motor activation. When TMS was applied 200 and 300 ms after the double consonant, the mere perception of /ll/ in Italian rare words evoked larger MEPs than /ll/ in frequent words. This finding suggests that the motor cortex not only participates in phonological encoding during speech perception but also interacts with top-down lexical processes. Note that parallels and shared representations between the perception and production systems have also been reported at the word level, with a similar temporal brain dynamic of phonological and lexicosemantic processes between the two modalities (Fairs et al., 2021; Strijkers & Costa, 2012; see also Strijkers et al., 2017). Additional evidence of speech-induced motor activity comes from D'Ausilio and colleagues (2014), who applied ultrasound tissue Doppler imaging (UTDI) to record 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/) elicited kinematic patterns that mirrored those 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 (see also Tang et al., 2021, for disrupted adaptation during production after the perception of altered vowel formants when receiving rTMS on the tongue motor area).

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, typically activated during speech production, when merely perceiving consonant–vowel (CV) syllables, compared to nonspeech sounds (white noise or bell rings). Pulvermüller and colleagues (2006) further supported the specificity of this motor activity, as consonants requiring different places of articulation for production (/p/ or /t/) activated the precentral gyrus somatotopically (lip or tongue motor area, respectively). 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 distinguish these same consonants from production-related activity in the (pre)motor cortex (i.e., pre/postcentral gyri and central sulcus) was unable to discriminate them when they were only perceived (see also Cheung et al., 2016, for somatotopy of place of articulation in speech production but not perception). Despite these discrepancies, these findings still suggest that articulatory features of perceived 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 or 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/). Results revealed that the classifier discriminated 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 and the right precentral and inferior frontal gyri. The temporoparietal 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 on place of articulation). Overall, these studies therefore support that in addition to temporoparietal 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 (Liberman et al., 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000; Schwartz et al., 2012).

FUNCTIONAL CONTRIBUTION OF THE MOTOR CORTEX TO SPEECH PERCEPTION

The role of the motor system as an essential or a subsidiary component of speech perception has, however, been hotly debated. The auditory system alone has been argued to be sufficient for decoding speech sounds through the analysis of spectrotemporal acoustic patterns, that is, by using general auditory mechanisms (see Diehl et al., 2004). The dual stream model (Hickok & Poeppel, 2007) also proposed that early speech processing begins bilaterally in auditory regions, to later split into two pathways. While a ventral stream in the temporal lobe would support speech comprehension (i.e., access to meaning), a dorsal stream would translate acoustic signals into articulatory representations in the premotor and inferior frontal cortices, via the parietotemporal junction, with a predominant role in speech acquisition and production. As to speech perception, it is proposed that the dorsal auditory-motor stream would participate only in sublexical processing such as syllable identification, that is, when listeners are required to specifically attend to phonemic information. As advocated by Hickok and his colleagues, motor resources would otherwise not be engaged in naturalistic listening conditions, thus only modestly contributing to speech perception (Hickok, 2012; Hickok & Poeppel, 2004, 2007; Stokes et al., 2019). This perspective is especially supported by clinical findings showing that patients with speech production deficits due to brain lesions in the motor system or the parietotemporal junction often retain intact speech comprehension (Hickok et al., 2011; Rogalsky et al., 2011). Such evidence has led to the interpretation that the neural substrates for speech production may function independently from those for perception.

Building on empirical neuroimaging findings (see below), other theories and models on the contrary support a functional contribution of the motor system to speech perception, emphasizing the sensorimotor nature of speech sounds rather than their purely auditory or motor character (Pulvermüller & Fadiga, 2010; Schomers & Pulvermüller, 2016; Schwartz et al., 2012). Accordingly, the auditory and motor systems would dynamically interact at all stages of speech perception, from acoustic–phonetic to phonemic processing, giving the motor system a primary role (Liebenthal & Möttönen, 2018). Motor regions would in particular instantiate internal forward predictions of the sensory input, influencing auditory processing and constraining phonemic categorization (Iacoboni, 2008; Rauschecker & Scott, 2009; see Grisoni & Pulvermüller, 2022, for a neural marker of phonological prediction in motor areas).

Evidence for tight reciprocal functional links between speech 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. Conversely, changes in the articulatory configuration can impact speech perception in adults (Ito et al., 2009) as well as in infants (Bruderer et al., 2015; Choi et al., 2019). At the age of 6 months, 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 to speech perception early in the course of language development.

TMS studies further provided positive evidence for a causal motor role in speech perception, as temporarily disrupting brain motor areas alters phonemic processing (see Möttönen & Watkins, 2012, for a review). Möttönen and Watkins (2009) found that the identification and discrimination of computer-generated CV syllables lying on a continuum and differing on their place of articulation (labials vs. dentals, e.g., /ba/–/da/ and /pa/–/ta/) were impaired when the lip representation in the left primary motor cortex was momentarily disrupted by repetitive TMS (rTMS). This impairment in categorical perception was not observed when either rTMS was applied over the left-hand motor representation or the perceived speech sounds did not involve the lips to be produced (e.g., /ka/–/ga/ and /da/–/ga/; see also Rogers et al., 2014). A follow-up study using TMS confirmed the motor contribution to the discrimination of speech sounds by showing that brain activity in the lip motor area reflects speech perception per se (i.e., sensitivity for between-category pairs along the continuum) and not post-perceptual decision making or response selection processes (Smalle et al., 2015). In line with this finding, Möttönen and colleagues (2013) reported an effect of motor cortical disruption by rTMS on brain responses to speech sounds, as recorded with electroencephalography (EEG), even in the absence of any behavioral task and when stimuli were not the focus of attention. More precisely, while participants watched a silent movie, temporarily disrupting their lip (but not hand) area in the left motor cortex reduced the early automatic mismatch negativity responses to phonetic changes (infrequent /ba/ or /ga/ in a sequence of /da/), but not to acoustic (duration) changes in speech and non-speech sounds (piano tones). This suggests that the motor system causally affects auditory speech processing when decoding articulatory features of phonemes. Surprisingly, however, the motor interference was not specific to the lip-related phoneme /b/ (relative to /g/), which seems at odds with previous work showing articulatory specific effects (Fadiga et al., 2002; Möttönen & Watkins, 2009). The authors suggested that such effects may depend on attention and particularly occur when participants need to focus on critical phonetic features that are task-relevant. A combined TMS-MEG (magnetoencephalography) study by the same research team corroborated this interpretation (Möttönen et al., 2014): rTMS over the left lip motor area affected the early left-lateralized brain responses to the labial consonant /b/, but not to velar or dental consonants (/g/ or /d/, respectively), when participants were required to respond to these stimuli. In contrast, when those same speech sounds were ignored, lip motor cortical disruption affected bilateral brain responses to the three consonants similarly and in a later time-window after their onset. This suggests that attention facilitates auditory-motor integration in the left hemisphere for processing specific articulatory features of speech sounds, but that left motor regions still automatically interact with bilateral auditory regions in an unspecific manner, that is, irrespective of place of articulation, at a later phonological stage.

Altogether, these findings highlight the close interactive mechanisms at play between the auditory and motor systems in the decoding of speech sounds (Liebenthal & Möttönen, 2018) and provide support for a causal role of motor regions in speech perception.

THE MOTOR SYSTEM'S FUNCTIONAL ROLE IN CHALLENGING SPEECH PERCEPTION

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 (Figure 1). D'Ausilio and colleagues (2009) revealed that priming activity of the lip motor area with TMS facilitated the recognition of labials (e.g., /p/) masked by white noise, while stimulating the tongue motor area enhanced the recognition of dentals (e.g., /t/). In their follow-up study (D'Ausilio et al., 2012), they in fact reported that such facilitation effects were only observed when syllables were embedded in noise, not when they were intact. Conversely, temporarily disrupting left

premotor cortex activity by 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 similar findings).

Converging evidence of specific motor activity during degraded speech processing comes from Callan and coworkers (2010) who highlighted the contribution of the ventral premotor cortex to distinguish between correct and incorrect phonemes in noisy conditions. 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 additionally 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 confirm that premotor regions selectively come into play when processing degraded though still identifiable speech sounds.

Du and colleagues (2014) reached similar conclusions using MVPA to examine the encoding of syllables that varied in articulatory features and were presented under different perceptual conditions. They showed that while activity in temporal regions exhibited good phoneme categorization only in the absence of noise, activity within premotor regions appeared more robust to distinguish degraded phonemes. These findings therefore suggest that the motor system can help speech processing, in moderately noisy conditions at least. They also align with the view that (pre)motor regions are part of a sensorimotor circuit that simulates articulatory gestures to anticipate sensory outcomes through the use of internal models (see Callan et al., 2004; Grisoni & Pulvermüller, 2022; Rauschecker & Scott, 2009; Skipper et al., 2007; Wilson et al., 2004; but see Hickok, 2012). Such top-down predictive coding likely enhances speech perception when the perceptual condition is challenging, underlining the role of the motor system in compensatory perceptual mechanisms that constrain and facilitate speech comprehension (Callan et al., 2004).

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) observed that this motor activity occurs irrespective of the nature of the degradation, either extrinsic or intrinsic to the speech signal. TMS applied over the lip motor region induced significantly larger MEPs during the perception of distorted than intact speech sounds, both when the distortion was caused by white noise masking and when it resulted from obstruction of the speaker's lip and tongue movements. Nuttall et al. 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 heard 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.

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 sections. 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 (Figure 1). 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, and nasals) belonging to different languages and varying in producibility for English native speakers. Compared 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, a region-of-interest 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 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. Echoing Wilson and Iacoboni's (2006) results in the temporal cortices, they 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 fMRI studies showing strong bilateral (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/, which are hardly distinguished by Japanese speakers, even with English experience, has indeed been shown to enhance activity in a bilateral network encompassing articulatory cortical regions in those participants (Callan et al., 2003). Interestingly, when native speakers of English performed the task on these same English phonemes but produced by Japanese speakers, therefore with a foreign accent, strong bilateral premotor involvement was also found (Callan et al., 2014). Converging findings were reported from phonetic training studies. A bilateral network including the inferior frontal gyrus, involved in articulatory processes, was activated after English native adults were trained for 5 hours to identify the Hindi dental retroflex consonant /ɖ/ (Golestani & Zatorre, 2004). This network was highly comparable to that recruited for the processing of the native consonants /d/ and /t/. On the other hand, Callan and colleagues (2004) showed an extension of activity in the premotor cortex, Broca's area, and supramarginal gyrus (among other cortical and subcortical regions)

from the left hemisphere to both hemispheres, after Japanese native speakers were extensively exposed for 1 month to the English difficult /ɹ/-/l/ contrast. Contrary to Golestani and Zatorre's findings, however, this pattern of activation spread far beyond that observed for the perception of an easy contrast (/b/-/g/, not trained) that also exists in Japanese. The authors interpreted their findings as reflecting the need to establish auditory–articulatory mappings when acquiring new phonemic categories, thus engaging additional neural resources also in the right hemisphere, to ease non-native speech acquisition and perception. Alternatively, the increased motor activation for non-native speech sounds observed across studies could reflect the recruitment of established native articulatory patterns, thus possibly interfering with foreign language processing and learning. Future neuromodulation studies targeting the motor cortex with TMS or transcranial direct current stimulation (tDCS) in pre- and post-training paradigms may help decipher if and when motor regions facilitate or hinder the establishment of non-native phonemic categories.

Altogether, previous findings support the idea that listeners recruit brain regions involved in speech production to process heard speech, especially under adverse auditory conditions (Figure 1). 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, 2017). Such a functional contribution raises the question of whether processing and learning 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.

CLASSICAL LEARNING PARADIGMS FOR FOREIGN PHONEMES: HIGH VARIABILITY PHONETIC TRAINING

Learning speech sounds that are not part of our phonological inventory is challenging, especially in adulthood. Since 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 non-native way of pronouncing foreign phonemes, a phenomenon commonly experienced as a foreign accent. The proximity between the phonological systems in the native and the foreign languages has been advocated as a major factor that influences learning new language's phonemes. According to Flege's speech learning model (Flege, 1995), foreign speech sounds perceived as close to native 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; 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 high variability phonetic training (HVPT; Logan et al., 1991), which is 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 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. 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 the 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. To assess the generalization of learning, both trained and new tokens, produced by the same or by different speakers, are usually included.

Numerous studies have shown improvement of learners' perceptual performance after 3–4 weeks of HVPT (15–45 training sessions), mostly regarding the English /ɹ/-/l/ contrast that Japanese native speakers 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 6 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 (Wang et al., 1999, 2003). For instance, in a classical HVPT paradigm varying consonantal contexts and speakers (Lambacher et al., 2005), Japanese native speakers exhibited higher identification of vowels from American English at post-test, in particular for those that were more distant from their native repertoire (/ɔ/ and /ɝ/). Interestingly, perceptual identification training has also proved successful on speech production, despite no explicit articulatory instruction being 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, Lambacher et al. (2005) also showed that, at post-test, the American English vowels produced by Japanese learners were better identified by native speakers, and their spectral overlap was reduced, compared to pre-test. This was especially true for more distant vowels (/ɔ/, /ɝ/ 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 greater improvements, both in perception and production, for 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 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 versus single talkers in HVPT. Whereas the meta-analysis by Zhang, Cheng, and Zhang (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, for lack of replication of Bradlow & Bent, 2008, on foreign-accented speech). Altogether, it 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 can also boost learning (Iverson et al., 2005; Ylinen et al., 2010; Zhang et al., 2009). Chinese native adults, for instance, learned 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 not being informative to distinguish the vowel categories (Cheng et al., 2019; see also Zhang, Cheng, 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 (e.g., speech-shaped noise, multitalker babble) also proved to benefit foreign phoneme identification (Cooke & Garcia Lecumberri, 2018; Leong et al., 2018). In the HVPT study by Mi and colleagues (2021), Chinese native speakers who learned English vowels embedded in a multitalker 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 3 months after training. Hence, adding background noise during training can help develop 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, Cheng, & Zhang, 2021). Given the functional role of motor regions in challenging speech perception, it is also possible, although this was not discussed by the authors, that training in noise may encourage 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).

AUDIOVISUAL TRAINING PARADIGMS FOR LEARNING FOREIGN PHONEMES

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 learned foreign phonemes (Hardison, 2003, 2005; Hazan et al., 2005; Inceoglu, 2016; Li & Somlak, 2019; Navarra & Soto-Faraco, 2007; Wang et al., 2014). Pereira Reyes and Hazan (2021) found comparable improvement in English vowel identification and production by Spanish native speakers following audiovisual and auditory phonetic training. Remarkably, training only with 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 studies, however, have 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) and which showed similar perceptual improvement after audiovisual and auditory training. Better pronunciation of this latter contrast was nevertheless observed after audiovisual than after auditory training, suggesting that information on articulatory gestures improves 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, and 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. 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.

EMBODIED TRAINING PARADIGMS FOR LEARNING FOREIGN PHONEMES

Spontaneous hand gestures usually 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 of speech and gestures arises early during native language development (Goldin-Meadow, 2010; Iverson, 2010), and gestures keep on easing language production 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 enhance 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; see Kühne & Gianelli, 2019, for a review). Performing iconic gestures while speaking can furthermore aid listeners' word comprehension under moderately adverse conditions, for instance, when the acoustic signal is spectrally degraded (Drijvers & Özyürek, 2017; Drijvers et al., 2018), both in young and in older adults (Schubotz et al., 2021). Interestingly, such benefit has also been reported in highly proficient non-native listeners, albeit to a lesser extent than in native listeners (Drijvers & Özyürek, 2018, 2020; Drijvers, Vaitonytė, & Özyürek, 2019; Drijvers, van der Plas, et al., 2019). It was proposed that a more intelligible auditory signal is required for non-native listeners to optimally map this information with the semantic information conveyed by the manual gestures, and thus to benefit from these extra cues. Although these studies examined word-level comprehension in high-proficiency non-natives, and despite the potential boosting effect of noise in learning foreign phonemes (e.g., Cheng et al., 2019; Mi et al., 2021), these findings deserve further attention for training paradigms combining speech with manual gestures to promote non-native speech sound learning. The last decade has indeed 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 et al., 2014; Li et al., 2020, 2021; Xi et al., 2020; Zheng et al., 2018).

Pitch

Several studies showed 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) 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, revealed 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). Notably, 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 supra-segmental 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 & Prieto, 2023; Baills et al., 2018; Gluhareva & Prieto, 2017). In Kushch's (2018) work, 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, Santiago, et al. (2022) found that Catalan learners improved in French accent in an oral reading task after training with sentence-level prosodic (pitch) imitated gestures (but see Baills, Alazard-Guiu, & Prieto, 2022, for contradictory findings).

Vowels

Besides prosodic patterns, embodied training paradigms have also been developed to encode segmental information such as vowel-length contrasts (see Table 1 for an overview of studies on gestural paradigms for foreign vowels). 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 nonreferential 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 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 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 (but see Hirata et al., 2014, who found similar results with an additional training condition based on the rhythmicity of Japanese moras using hand flicks only, one for short vowels and two for long vowels). The potential lack of obvious

Table 1. Overview of studies using gestural learning paradigms for foreign vowels

Study	Native/ foreign language	Foreign phonetic features & phonemes	Training[1]			Testing task/measurement		Improvement			
								Perception		Production	
			Gesture	Experimental groups	Duration	Perception	Production	Post-test	Follow-up	Post-test	Follow-up
Hirata & Kelly, 2010	English/ Japanese	vowel length /a/ vs. /aː/ /i/ vs. /iː/ /u/ vs. /uː/ /e/ vs. /eː/ /o/ vs. /oː/	beat (flick) vs. duration (sweep)	(1) audio (2) audiovisual (3) audio + gestural (4) audiovisual + gestural	4 sessions (30 min/ session) over 2 weeks	identification	N/A	all groups; better in audiovisual group	N/A	N/A	
Hirata et al., 2014	English/ Japanese	vowel length /a/ vs. /aː/ /u/ vs. /uː/ /o/ vs. /oː/	short (one flick) vs. long (syllable: long dip + flick, or mora: 2 flicks)	(1) audiovisual + syllable gestural observation (2) audiovisual + syllable gestural imitation (3) audiovisual + mora gestural observation (4) audiovisual + mora gestural imitation	4 sessions (20–30 min/ session) on 2 non-consecutive days	identification	N/A	all groups	N/A	N/A	
Li et al., 2020	Catalan/ Japanese	vowel length /e/ vs. /eː/ /o/ vs. /oː/	duration (sweep): short vs. long	(1) audiovisual (2) audiovisual + gestural	one session (2.5 min)	identification	acoustic analysis: mean duration	both groups	N/A	gestural group	N/A
Li et al., 2023	Catalan/ French	front rounded vowels /y, ø, œ/	prosodic	(1) audiovisual (2) audiovisual + gestural	3 sessions (15 min/ session) over 3 weeks	N/A	perceptual rating; acoustic analysis: formants	N/A		gestural group	2 weeks later; gestural group
Xi et al., 2024	Catalan, Spanish/ English	vowel height and backness /æ, ʌ/	iconic	(1) audiovisual (2) audiovisual + lip gestural (3) audiovisual + tongue gestural	one session (45 min)	identification	acoustic analysis: formants	all groups	1 week later; no group	lip gestural group	1 week later; lip gestural group only for /æ/
Hoetjes & van Maastricht, 2020	Dutch/ Spanish	rounded vowel /u/	iconic vs. pointing	(1) audio (2) audiovisual (3) audiovisual + pointing gestural (4) audiovisual + iconic gestural	one session (3–4 min)	N/A	perceptual rating	N/A		all audiovisual groups	N/A

[1] Learners were required to repeat phonemes during training only in Li et al. (2020, 2023) studies.

Green cells indicate that gestural training improved learning of foreign speech sounds, whereas yellow and red cells stand for, respectively, limited and lack of improvement. N/A = not tested in the study; min = minute(s).

correspondence between the hand flick gesture and the short vowel for English native listeners, 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 these two studies. Li and colleagues (2020) on the other hand reported 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/) in Catalan adults. This advantage of durational gestures over training without gestures was, however, found only for non-native phoneme production (see also Li et al., 2023, for effects of prosodic gestures on the production of French front-rounded vowels by Catalan speakers). Identification performance on the contrary improved similarly following the two types of training. Hence, in line with the previously mentioned work (Hirata & Kelly, 2010; Hirata et al., 2014), hand gestures failed to improve the perception of vowel-length contrasts, whether these gestures were merely observed or reproduced. This may be explained by the major challenge in acquiring this type of contrast compared to pitch contrasts for non-native speakers (Hirata, 2015). Despite the encouraging results on the production side at least, 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 needed to target the right gestures and fully support the beneficial role of hand gestures on the learning of non-native durational vowel contrasts.

Phonetic Features of Consonants

What about phonetic features? Can manual gestures that explicitly code for place or manner of articulation help with learning non-native speech sounds? A handful of recent studies have tackled this issue, with generally promising results, in particular on speech sound production (e.g., Amand & Touhami, 2016, for unreleased plosives; Ozakin et al., 2023, for fricatives; Xi et al., 2023, for vowel lip aperture; see Table 2 for an overview of studies targeting consonants with gestural learning paradigms). 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 gesture. 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 5-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. In contrast, identification performance for both plosives and affricates did not benefit from hand gestures compared to the no-gesture training condition (in line with previous work on vowels; Hirata & Kelly, 2010; Hirata et al., 2014; Li et al., 2020). These findings emphasize that only manual gestures that appropriately reflect the phonetic features of non-native phonemes may foster their acquisition in adults and improve their pronunciation. A follow-up study (Li et al., 2021) confirmed the importance not only of the addition of manual 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 more 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

Table 2. Overview of studies using gestural learning paradigms for foreign consonants

Study	Native/ foreign language	Foreign phonetic features & phonemes	Training[1]			Testing task/measurement		Improvement			
								Perception		Production	
			Gesture	Experimental groups	Duration	Perception	Production	Post-test	Follow-up	Post-test	Follow-up
Hoetjes & van Maastricht, 2020	Dutch/ Spanish	labiodental /θ/	iconic vs. pointing	(1) audio (2) audiovisual (3) audiovisual + pointing gestural (4) audiovisual + iconic gestural	one session (3–4 min)	N/A	perceptual rating	N/A		pointing gestural group	N/A
Xi et al., 2020	Catalan/ Mandarin	aspirated plosives /p/ vs. /p^h/ /t/ vs. /t^h/ /k/ vs. /k^h/	fist-to-open	(1) audio visual (2) audiovisual + gestural	one session (5 min 36 s)	identification	perceptual rating	both groups	N/A	gestural group	N/A
		aspirated affricates /ts/vs./ts^h/ /tɕ/vs./$tɕ^h$/ /tʂ/vs./$tʂ^h$/								no group	
Li et al., 2021	Catalan/ Mandarin	aspirated plosives /p/ vs. /p^h/ /t/ vs. /t^h/ /k/ vs. /k^h/	fist-to-open	(1) audiovisual (2) audiovisual + gestural	one session (5 min)	identification	gestural rating; acoustic analysis: VOT	both groups	N/A	audiovisual and well-performed gestural groups	3 days later; well-performed gestural group

[1] Learners were required to repeat phonemes during training only in Li et al. (2021).

Green cells indicate that gestural training improved learning of foreign speech sounds, whereas yellow and red cells stand for, respectively, limited and lack of improvement. N/A = not tested in the study; min = minute(s); s = second(s); VOT = voice onset time.

training) as well as 3 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 nonvisible articulatory features also seem to impact learning efficiency. In the study by Hoetjes and van Maastricht (2020), Spanish adults learned to produce 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 (simpler) 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, for example, to acquire a non-native phoneme outside the native phonological inventory, providing complex hand gestures may be detrimental to learning (see Kelly & Lee, 2012, for similar arguments). Notably, 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 anteroposterior 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 (e.g., Hirata et al., 2014; Li et al., 2020, 2021; Xi et al., 2020). For production, however, results revealed that the lip-related gesture helped the learners to adjust their lip aperture (as measured by formant values) for non-native vowels more than the tongue-related gesture and nongestural 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 nonvisible features, involving the tongue in particular, as the latter may potentially increase the processing demands. Indeed, 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), might also explain the limited learning advantage of tongue-related hand gestures. In 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 participants imitated hand

gestures that illustrated the shape of the tongue. The fact that learners only observed the manual gestures in Xi et al. (2024) and Hoetjes and van Maastricht (2020) suggests that actually performing the gestures may be a key ingredient for efficient learning. Although this interpretation is in line with most studies on vocabulary learning or more general cognitive skills (e.g., Goldin-Meadow et al., 2009; Macedonia et al., 2011; but see Kelly et al., 2014), the few studies that compared observation and imitation for foreign phoneme learning have provided mixed results (advantage of imitation over observation: e.g., Baills et al., 2019; Kushch, 2018; no advantage: e.g., Hirata et al., 2014; Kelly et al., 2014). While manual gestures illustrating nonvisible articulatory features may be difficult to integrate with incongruent visual facial cues, they might still be effective when actively and correctly imitated, 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 the various gestures employed and the few phonemes investigated across different languages. Yet, there are some implications for future training paradigms to build new phonemic categories so as to improve both perception and production in a non-native language. Manual gestures that emphasize distinctions between phonemes should precisely represent the articulatory features of the target foreign speech sounds to acquire. Some challenges still remain though, for instance, for those articulatory features that are not directly visible (e.g., tongue position or shape) and that did not benefit so well from manual gestures during training, at least when these were merely observed, or on the perception side. Assessing the motor performance of learners for these gestures during the training phase could also be crucial to maximize learning efficacy. In addition, longer training paradigms, currently absent in the literature to the best of our knowledge, may be beneficial in strengthening the link between manual gestures and perceived articulatory features, thereby further improving the perception and production of foreign phonemes.

CONCLUSION

We 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 embodied multisensory information. Current neuroimaging evidence indicates that the (pre)motor regions underlying speech production are also engaged in speech perception, underscoring the sensorimotor foundation of speech. Somatotopic 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 involvement in decoding articulatory features of perceived phonemes, the potential benefits of sensorimotor-based training for learning become evident. Whereas training paradigms that introduce high phonemic variability can enhance perceptual and production skills in foreign languages, multisensory learning protocols, such as training with manual gestures, appear as a promising venue to bolster the acquisition of non-native speech sounds. The limited number of studies, together with the effects restricted to production or to certain phonetic features, however, warrant further scrutiny to fully attest the benefits of gestural learning. Future research should implement longitudinal paradigms to assess whether both the perception and production of foreign speech sounds 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 of non-native speech sounds within the motor cortex. Studies using TMS or tDCS to stimulate motor cortex excitability may also provide further evidence to decide on the causal role of these brain regions in non-native

language acquisition and processing. This, in turn, will allow the refinement of training protocols to optimize the learning process as well as provide valuable insights into the role of embodied experiences during learning and development.

FUNDING INFORMATION

Tzuyi Tseng, LabEx ASLAN (https://dx.doi.org/10.13039/501100011602), Award ID: ANR-10-LABX-0081; ANR-11-IDEX-0007. Jennifer Krzonowski, LabEx ASLAN (https://dx.doi.org/10.13039/501100011602), Award ID: ANR-10-LABX-0081; ANR-11-IDEX-0007. Claudio Brozzoli, Agence Nationale de la Recherche (https://dx.doi.org/10.13039/501100001665), Award ID: ANR-19-CE28-0015. Claudio Brozzoli, Appel à Projets Pluridisciplinaires Interne (APPI), Université Lumière Lyon 2. Claudio Brozzoli, James S. McDonnell Foundation (https://dx.doi.org/10.13039/100000913), Award ID: doi.org/10.37717/2021-3101. Alice C. Roy, Agence Nationale de la Recherche (https://dx.doi.org/10.13039/501100001665), Award ID: ANR-19-CE28-0015. Alice C. Roy, Appel à Projets Pluridisciplinaires Interne (APPI), Université Lumière Lyon 2. Alice C. Roy, LabEx ASLAN (https://dx.doi.org/10.13039/501100011602), Award ID: ANR-10-LABX-0081; ANR-11-IDEX-0007. Véronique Boulenger, Agence Nationale de la Recherche (https://dx.doi.org/10.13039/501100001665), Award ID: ANR-19-CE28-0015. Véronique Boulenger, Appel à Projets Pluridisciplinaires Interne (APPI), Université Lumière Lyon 2. Véronique Boulenger, LabEx ASLAN (https://dx.doi.org/10.13039/501100011602), Award ID: ANR-10-LABX-0081; ANR-11-IDEX-0007.

AUTHOR CONTRIBUTIONS

Tzuyi Tseng: Conceptualization; Visualization; Writing – original draft; Writing – review & editing. **Jennifer Krzonowski**: Conceptualization. **Claudio Brozzoli**: Conceptualization; Funding acquisition; Writing – review & editing. **Alice C. Roy**: Conceptualization; Funding acquisition; Project administration; Writing – review & editing. **Véronique Boulenger**: Conceptualization; Funding acquisition; Project administration; Supervision; Writing – original draft; Writing – review & editing.

DATA AND CODE AVAILABILITY STATEMENTS

This study did not generate any new data or code.

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, PubMed: 19632341

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, PubMed: 29981784

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*, Article 394. https://doi.org/10.3389/fpsyg.2014.00394, PubMed: 24834062

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), 377–388. https://doi.org/10.1515/rela-2016-0020

Archila-Meléndez, M. E., Valente, G., Correia, J. M., Rouhl, R. P. W., van Kranen-Mastenbroek, V. H., & 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), Article ENEURO.0252-17.2018. https://doi.org/10.1523/ENEURO.0252-17.2018, PubMed: 29610768

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, PubMed: 26715582

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 *Proceedings of the 11th International Conference on Speech Prosody* (pp. 802–806). 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. In *Proceedings of the 9th International Conference on Speech Prosody* (pp. 853–857). ISCA. 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, PubMed: 24046079

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). MIT Press. https://doi.org/10.7551/mitpress/2387.003.0011

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. *Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378, PubMed: 11248981

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911, PubMed: 10499009

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, PubMed: 17532315

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. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310. https://doi.org/10.1121/1.418276, PubMed: 9104031

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*, Article 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, PubMed: 26460030

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, PubMed: 20184959

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*, Article 275. https://doi.org/10.3389/fnins.2014.00275, PubMed: 25232302

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, PubMed: 15219590

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, PubMed: 12781731

Cebrian, J., & Carlet, A. (2014). Second-language learners' identification of target-language phonemes: A short-term phonetic training study. *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*, Article 1178. https://doi.org/10.3389/fpsyg.2019.01178, PubMed: 31178795

Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in human motor cortex. *eLife*, *5*, Article e12577. https://doi.org/10.7554/eLife.12577, PubMed: 26943778

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, PubMed: 31037603

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*, Article 323. https://doi.org/10.3389/fnhum.2020.00323, PubMed: 32903691

Cooke, M., & Garcia Lecumberri, M. L. (2018). Effects of exposure to noise during perceptual training of non-native language sounds. *Journal of the Acoustical Society of America*, *143*(5), 2602–2610. https://doi.org/10.1121/1.5035080, PubMed: 29857707

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, PubMed: 26558773

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, PubMed: 21676385

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*, *369*(1644), Article 20130418. https://doi.org/10.1098/rstb.2013.0418, PubMed: 24778384

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, PubMed: 19217297

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179. https://doi.org/10.1146/annurev.psych.55.090902.142028, PubMed: 14744213

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

Drijvers, L., & Özyürek, A. (2017). Visual context enhanced: The joint contribution of iconic gestures and visible speech to degraded speech comprehension. *Journal of Speech, Language, and Hearing Research*, *60*(1), 212–222. https://doi.org/10.1044/2016_JSLHR-H-16-0101, PubMed: 27960196

Drijvers, L., & Özyürek, A. (2018). Native language status of the listener modulates the neural integration of speech and iconic gestures in clear and adverse listening conditions. *Brain and Language*, *177–178*, 7–17. https://doi.org/10.1016/j.bandl.2018.01.003, PubMed: 29421272

Drijvers, L., & Özyürek, A. (2020). Non-native listeners benefit less from gestures and visible speech than native listeners during degraded speech comprehension. *Language and Speech*, *63*(2), 209–220. https://doi.org/10.1177/0023830919831311, PubMed: 30795715

Drijvers, L., Özyürek, A., & Jensen, O. (2018). Hearing and seeing meaning in noise: Alpha, beta, and gamma oscillations predict gestural enhancement of degraded speech comprehension. *Human Brain Mapping*, *39*(5), 2075–2087. https://doi.org/10.1002/hbm.23987, PubMed: 29380945

Drijvers, L., Vaitonytė, J., & Özyürek, A. (2019). Degree of language experience modulates visual attention to visible speech and iconic gestures during clear and degraded speech comprehension. *Cognitive Science*, *43*(10), Article e12789. https://doi.org/10.1111/cogs.12789, PubMed: 31621126

Drijvers, L., van der Plas, M., Özyürek, A., & Jensen, O. (2019). Native and non-native listeners show similar yet distinct oscillatory dynamics when using gestures to access speech in noise. *NeuroImage*, *194*, 55–67. https://doi.org/10.1016/j.neuroimage.2019.03.032, PubMed: 30905837

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, PubMed: 24778251

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*, Article 12241. https://doi.org/10.1038/ncomms12241, PubMed: 27483187

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: Evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*(12), 4772–4788. https://doi.org/10.1093/cercor/bhv136, PubMed: 26157026

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, PubMed: 11849307

Fairs, A., Michelas, A., Dufour, S., & Strijkers, K. (2021). The same ultra-rapid parallel brain dynamics underpin the production and perception of speech. *Cerebral Cortex Communications*, *2*(3), Article tgab040. https://doi.org/10.1093/texcom/tgab040, PubMed: 34296185

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, PubMed: 18470815

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

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, PubMed: 36416451

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, *21*(5), 609–631. 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), 239–255. https://doi.org/10.3917/enf1.103.0239, PubMed: 23526836

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Review of Psychology*, *64*, 257–283. https://doi.org/10.1146/annurev-psych-113011-143802, PubMed: 22830562

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*(3), 267–272. https://doi.org/10.1111/j.1467-9280.2009.02297.x, PubMed: 19222810

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, PubMed: 14980552

Golestani, N., & Zatorre, R. J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2–3), 55–67. https://doi.org/10.1016/j.bandl.2008.01.005, PubMed: 18295875

Grauwinkel, K., Dewitt, B., & Fagel, S. (2007). Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (pp. 706–709). ISCA. https://doi.org/10.21437/Interspeech.2007-295

Grisoni, L., & Pulvermüller, F. (2022). Predictive and perceptual phonemic processing in articulatory motor areas: A prediction potential & mismatch negativity study. *Cortex*, *155*, 357–372. https://doi.org/10.1016/j.cortex.2022.06.017, PubMed: 36095883

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics in Language Teaching*, *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. *Journal of the Acoustical Society of America*, *140*(S4), Article 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*, Article 2051. https://doi.org/10.3389/fpsyg.2017.02051, PubMed: 29255435

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 contrasts. *Journal of the Acoustical Society of America*, *119*(3), 1740–1751. https://doi.org/10.1121/1.2166611, PubMed: 16583916

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, A. G., Carlyon, R. P., Johnsrude, I. S., & Davis, M. H. (2012). Brain regions recruited for the effortful comprehension of noise-vocoded words. *Language and Cognitive Processes*, *27*(7–8), 1145–1166. https://doi.org/10.1080/01690965.2012.662280

Hickok, G. (2012). The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *Journal of Communication Disorders*, *45*(6), 393–402. https://doi.org/10.1016/j.jcomdis.2012.06.004, PubMed: 22766458

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2011.01.019, PubMed: 21315253

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1–2), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011, PubMed: 15037127

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113, PubMed: 17431404

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*, Article 118577. https://doi.org/10.1016/j.neuroimage.2021.118577, PubMed: 34525395

Hirata, Y. (2015). L2 phonetics and phonology. In H. Kubozono (Ed.), *Handbook of Japanese phonetics and phonology* (pp. 719–762). De Gruyter Mouton. https://doi.org/10.1515/9781614511984.719

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), PubMed: 20220023

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (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, PubMed: 25088127

Hoetjes, M., & van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, *11*, Article 575032. https://doi.org/10.3389/fpsyg.2020.575032, PubMed: 33329219

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, PubMed: 18412134

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, PubMed: 18440208

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, PubMed: 19164569

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, PubMed: 20096145

Iverson, J. M., & Goldin-Meadow, S. (1998). Why people gesture when they speak. *Nature*, *396*(6708), Article 228. https://doi.org/10.1038/24300, PubMed: 9834030

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. *Journal of the Acoustical Society of America*, *122*(5), 2842–2854. https://doi.org/10.1121/1.2783198, PubMed: 18189574

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *Journal of the Acoustical Society of America*, *126*(2), 866–877. https://doi.org/10.1121/1.3148196, PubMed: 19640051

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. *Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307, PubMed: 16334698

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. [D.], 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), Article 7. https://doi.org/10.1525/collabra.76

Kelly, S. D., & Hirata, Y. (2017). What neural measures reveal about foreign language learning of Japanese vowel length contrasts with hand gestures. In S. Tanaka (Ed.), *New development in*

phonology research: Festschrift in honor of Haruo Kubozono (pp. 278–294). Kaitakusha.

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*, Article 673. https://doi.org/10.3389/fpsyg.2014.00673, PubMed: 25071646

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

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*, Article 108. https://doi.org/10.3389/fpsyg.2019.00108, PubMed: 30787892

Kushch, O. (2018). *Beat gestures and prosodic prominence: Impact on learning* [Ph.D. dissertation, Universitat Pompeu Fabra]. 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*, *4*(9), 29–34. https://doi.org/10.4236/ce.2013.49B007

Leong, C. X. R., Price, J. M., Pitchford, N. J., & van Heuven, W. J. B. (2018). High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLOS One*, *13*(10), Article e0204888. https://doi.org/10.1371/journal.pone.0204888, PubMed: 30300372

Li, P., Baills, 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., Baills, 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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279, PubMed: 4170865

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6, PubMed: 4075760

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*(5), 187–196. https://doi.org/10.1016/S1364-6613(00)01471-6, PubMed: 10782105

Liebenthal, E., & Möttönen, R. (2018). An interactive model of auditory-motor speech perception. *Brain and Language*, *187*, 33–40. https://doi.org/10.1016/j.bandl.2017.12.004, PubMed: 29268943

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. *Journal of the Acoustical Society of America*, *94*(3), 1242–1255. https://doi.org/10.1121/1.408177, PubMed: 8408964

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. *Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149, PubMed: 7963022

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649, PubMed: 2016438

Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. *Frontiers in Psychology*, *5*, Article 1467. https://doi.org/10.3389/fpsyg.2014.01467, PubMed: 25538671

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, PubMed: 20645312

Massaro, D. W. & Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of the 8th European Conference on Speech Communication and Technology* (pp. 2249–2252). ISCA. https://doi.org/10.21437/Eurospeech.2003-629

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*(4–5), 657–662. https://doi.org/10.1016/S0031-9384(02)00916-2, PubMed: 12527015

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0, PubMed: 1012311

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (Ed.). (2000). *Language and gesture*. Cambridge University Press. https://doi.org/10.1017/CBO9780511620850

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, PubMed: 17900904

Mi, L., Tao, S., Wang, W., Dong, Q., Dong, B., Li, M., & Liu, C. (2021). Training non-native vowel perception: In quiet or noise. *Journal of the Acoustical Society of America*, *149*(6), 4607–4619. https://doi.org/10.1121/10.0005276, PubMed: 34241439

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, PubMed: 35427446

Morett, L. M. (2023). Observing gesture at learning enhances subsequent phonological and semantic processing of L2 words: An

N400 study. *Brain and Language*, *246*, Article 105327. https://doi.org/10.1016/j.bandl.2023.105327, PubMed: 37804717

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, PubMed: 19657034

Möttönen, R., & Watkins, K. E. (2012). Using TMS to study the role of the articulatory motor system in speech perception. *Aphasiology*, *26*(9), 1103–1118. https://doi.org/10.1080/02687038.2011.619515, PubMed: 22942513

Möttönen, R., Dutton, R., & Watkins, K. E. (2013). Auditory-motor processing of speech sounds. *Cerebral Cortex*, *23*(5), 1190–1197. https://doi.org/10.1093/cercor/bhs110, PubMed: 22581846

Möttönen, R., van de Ven, G. M., & Watkins, K. E. (2014). Attention fine-tunes auditory-motor processing of speech sounds. *Journal of Neuroscience*, *34*(11), 4064–4069. https://doi.org/10.1523/JNEUROSCI.2214-13.2014, PubMed: 24623783

Murakami, T., Kell, C. A., Restle, J., Ugawa, Y., & Ziemann, U. (2015). Left dorsal speech stream components and their contribution to phonological processing. *Journal of Neuroscience*, *35*(4), 1411–1422. https://doi.org/10.1523/JNEUROSCI.0246-14.2015, PubMed: 25632119

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*(1), 4–12. https://doi.org/10.1007/s00426-005-0031-5, PubMed: 16362332

Nishi, K., & Kewley-Port, D. (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), PubMed: 18055770

Nishi, K., & Kewley-Port, D. (2008). Nonnative speech perception training using vowel subsets: Effects of vowels in sets and order of training. *Journal of Speech, Language, and Hearing Research*, *51*(6), 1480–1493. https://doi.org/10.1044/1092-4388(2008/07-0109), PubMed: 18664694

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, PubMed: 27884757

Nuttall, H. E., Kennedy-Higgins, D., Devlin, J. T., & Adank, P. (2018). Modulation of intra-and inter-hemispheric connectivity between primary and premotor cortex during speech perception. *Brain and Language*, *187*, 74–82. https://doi.org/10.1016/j.bandl.2017.12.002, PubMed: 29397191

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, PubMed: 26732405

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. In D. W. Massaro, J. Light, & K. Geraci (Eds.), *International Conference on Auditory-Visual Speech Processing (AVSP 2001)* (pp. 149–154). Auditory-Visual Speech Association. https://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, PubMed: 20932914

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

Pereira Reyes, Y., & Hazan, V. (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. *Onomázein*, *51*, 111–136. https://doi.org/10.7764/onomazein.51.04

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. *Journal of the Acoustical Society of America*, *130*(1), 461–472. https://doi.org/10.1121/1.3593366, PubMed: 21786912

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. *Journal of the Acoustical Society of America*, *119*(3), 1684–1696. https://doi.org/10.1121/1.2161427, PubMed: 16583912

Pulvermüller, F., & Fadiga, L. (2010). Active perception: Sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience*, *11*(5), 351–360. https://doi.org/10.1038/nrn2811, PubMed: 20383203

Pulvermüller, 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, PubMed: 16682637

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, PubMed: 19471271

Rogalsky, C., Love, T., Driscoll, D., Anderson, S. W., & Hickok, G. (2011). Are mirror neurons the basis of speech perception? Evidence from five cases with damage to the purported human mirror system. *Neurocase*, *17*(2), 178–187. https://doi.org/10.1080/13554794.2010.509318, PubMed: 21207313

Rogers, J. C., Möttönen, R., Boyles, R., & Watkins, K. E. (2014). Discrimination of speech and non-speech sounds following theta-burst stimulation of the motor cortex. *Frontiers in Psychology*, *5*, Article 754. https://doi.org/10.3389/fpsyg.2014.00754, PubMed: 25076928

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, PubMed: 18440210

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, PubMed: 29114768

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*, Article 1318. https://doi.org/10.3389/fpsyg.2014.01318, PubMed: 25505434

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, PubMed: 19362734

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, PubMed: 29317325

Schomers, M. R., & Pulvermüller, F. (2016). Is the sensorimotor cortex relevant for speech perception and understanding? An integrative review. *Frontiers in Human Neuroscience*, *10*, Article 435. https://doi.org/10.3389/fnhum.2016.00435, PubMed: 27708566

Schubotz, L., Holler, J., Drijvers, L., & Özyürek, A. (2021). Aging and working memory modulate the ability to benefit from visible speech and iconic gestures during speech-in-noise comprehension. *Psychological Research*, *85*(5), 1997–2011. https://doi.org/10.1007/s00426-020-01363-8, PubMed: 32627053

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

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, PubMed: 27821280

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, PubMed: 17218482

Smalle, E. H., Rogers, J., & Möttönen, R. (2015). Dissociating contributions of the motor cortex to speech perception and response bias by using transcranial magnetic stimulation. *Cerebral Cortex*, *25*(10), 3690–3698. https://doi.org/10.1093/cercor/bhu218, PubMed: 25274987

Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *TESOL Quarterly*, *51*(1), 59–89. https://doi.org/10.1002/tesq.276

Stokes, R. C., Venezia, J. H., & Hickok, G. (2019). The motor system's [modest] contribution to speech perception. *Psychonomic Bulletin & Review*, *26*(4), 1354–1366. https://doi.org/10.3758/s13423-019-01580-2, PubMed: 30945170

Strijkers, K., & Costa, A. (2012). The neurocognition of language production: Introduction to the special topic. *Frontiers in Psychology*, *3*, Article 198. https://doi.org/10.3389/fpsyg.2012.00198, PubMed: 22557944

Strijkers, K., Costa, A., & Pulvermüller, F. (2017). The cortical dynamics of speaking: Lexical and phonological knowledge simultaneously recruit the frontal and temporal cortex within 200 ms. *NeuroImage*, *163*, 206–219. https://doi.org/10.1016/j.neuroimage.2017.09.041, PubMed: 28943413

Tang, D.-L., McDaniel, A., & Watkins, K. E. (2021). Disruption of speech motor adaptation with repetitive transcranial magnetic stimulation of the articulatory representation in primary motor cortex. *Cortex*, *145*, 115–130. https://doi.org/10.1016/j.cortex.2021.09.008, PubMed: 34717269

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, PubMed: 21910556

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). On the use of an articulatory talking head for second language pronunciation training: The case of Chinese learners of French. In *Proceedings of 10th International Seminar on Speech Production* (pp. 449–452). MPG-PuRe.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *Journal of the Acoustical Society of America*, *113*(2), 1033–1043. https://doi.org/10.1121/1.1531176, PubMed: 12597196

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217, PubMed: 10615703

Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989–994. https://doi.org/10.1016/S0028-3932(02)00316-0, PubMed: 12667534

Werker, J. F., Frost, P. E., & McGurk, 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, PubMed: 1286433

Wik, P., & Engwall, O. (2008). Looking at tongues—Can it help in speech perception? In *Proceedings FONETIK 2008* (pp. 57–60). Universität zu Köln. 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, PubMed: 16919478

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702. https://doi.org/10.1038/nn1263, PubMed: 15184903

Xi, X., Li, P., Bailly, 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, PubMed: 33090915

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). Université Grenoble-Alpes. 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*, *74*(4), 884–916. 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, PubMed: 34370501

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, PubMed: 19445609

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, PubMed: 20080724

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*, Article 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, PubMed: 34763529

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, PubMed: 19457395

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, PubMed: 30877849

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, PubMed: 30193334