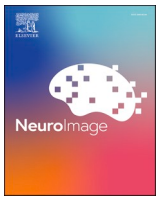
NeuroImage 310 (2025) 121169

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



Motor-based prediction mediates implicit vocal imitation

Yuchunzi Wu [a,b,c,*], Zhili Han [d], Xing Tian [a,b,c,*]

[a] Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning; Division of Arts and Sciences, New York University Shanghai, Shanghai, China
[b] NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China
[c] Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China
[d] NingboTech University, Ningbo, Zhejiang, China

ARTICLE INFO

Keywords:
phonetic convergence
imitation
prediction
forward model
mismatch negativity

ABSTRACT

Phonetic convergence—the unconscious adaptation of one's speech to resemble that of an interlocutor—is thought to arise from predictive mechanisms. Two types of predictions have been proposed to modulate others' speech: memory-based predictions, which reduce sensitivity to acoustic features reflecting a speaker's vocal identity, and motor-based predictions, which are grounded in the listener's own vocal characteristics. Compared to a relatively well-established role of memory-based predictions, whether motor-based predictions suppress or enhance sensitivity to listener-matched predicted features and how they contribute to phonetic convergence remain unclear. In the present study, we examined these processes using a novel speaking oddball task in which participants were randomly prompted to repeat words they heard. Auditory mismatch negativity served as a neural index of mismatch detection. Prior to the oddball task, participants were divided into a shadow group—engaging in an additional shadowing task to promote vocal convergence—and a non-shadow group that did not receive such exposure. EEG analyses revealed that motor-based predictions enhance sensitivity to listener-matched predicted features following convergence behaviour, with this enhancement correlating with greater vocal convergence. Our novel oddball design provided an efficient method for revealing the dynamic interplay between internal predictive signals and external inputs that mediates phonetic convergence. These findings challenge the view that motor-based predictions only suppress neural responses to predicted features, and instead highlight their potential role in enhancing perceptual learning and guiding vocal adjustments. Motor-based predictions orchestrate sensorimotor interaction and memory-based operations to mediate implicit learning behaviour in a social context.

1. Introduction

Adaptation is a fundamental behaviour that allows living organisms to navigate dynamic environments, and in humans, this adaptability extends into social interactions. A notable example is phonetic convergence, also known as vocal mimicry, where individuals subconsciously adopt the acoustic characteristics of speech they hear, leading to increased vocal resemblance with their interlocutors (Gambi and Pickering, 2013; Pardo et al., 2017). This phenomenon occurs across diverse languages and dialects, both in naturalistic social interactions and controlled experimental settings (e.g., Babel and Bulatov, 2012; Bosshardt et al., 1997; Brouwer et al., 2010; Delvaux and Soquet, 2007; Olmstead et al., 2013, 2021; Pardo, 2006; Pardo et al., 2013; Wagner et al., 2021). Phonetic convergence typically involves imitating sub-phonemic features—acoustic variations that do not alter word meanings but convey a speaker's physical or social traits, such as gender, age, or cultural-geographic background. This distinction underscores that phonetic convergence targets paralinguistic properties of speech—particularly identity cues (e.g., voice timbre and speech patterns shaped by one's anatomical traits and social environment)—rather than phonological elements that differentiate between distinct phonemes (e.g., /b/ versus /d/ in bark and dark). Convergence behaviour is thought to play a crucial role in fostering social cohesion and building rapport (Dragojevic et al., 2015; Giles et al., 1991), while also enhancing communication by streamlining comprehension and enabling individuals to predict their interlocutors' speech (Adank et al., 2010; Pickering and Gambi, 2018). Nevertheless, the specific neural and cognitive mechanisms that drive this adaptive behaviour remain

* Correspondence authors: Division of Arts and Sciences, New York University Shanghai, Shanghai, China.
E-mail addresses: yw2062@nyu.edu (Y. Wu), xing.tian@nyu.edu (X. Tian).

https://doi.org/10.1016/j.neuroimage.2025.121169
Received 4 December 2024; Received in revised form 9 March 2025; Accepted 21 March 2025
Available online 22 March 2025

unclear.

The integrated theory of language production and comprehension posits that we continuously predict upcoming speech both when we speak and when we listen to others, with phonetic convergence emerging from the minimization of prediction errors in others' speech (Gambi and Pickering, 2013; Pickering and Garrod, 2013). Whilst speaking, the brain generates motor commands and develops internal forward models to anticipate the sensory outcomes of these commands (Hickok, 2012). When there is a mismatch between predicted and actual feedback, prediction errors emerge, prompting adjustments in both motor commands and forward models to ensure accurate and fluent speech. Evidence shows that auditory responses to self-generated speech are suppressed compared to responses to playback of the same speech signal (Curio et al., 2000; Greenlee et al., 2011; Ventura et al., 2009; Whitford, 2019). This speaking-induced suppression reflects the brain's signal cancellation of feedback that matches the prediction and is attenuated or abolished when the feedback is altered or replaced by an alien voice (Behroozmand and Larson, 2011; Heinks-Maldonado et al., 2007; Houde et al., 2002). Additionally, decreased suppression has been observed for less prototypical but natural utterances and is associated with greater corrective changes within those utterances (Niziolek et al., 2013). The authors interpreted these findings as evidence that forward model predictions represent sensory goals that do not encode phonetic variations resulting from movement variability. It is also possible that motor-based forward model predictions are sensitive to acoustic information reflecting a speaker's typical vocal traits—an integral component of his/her identity cues—which allows for the rapid detection of less prototypical utterances.

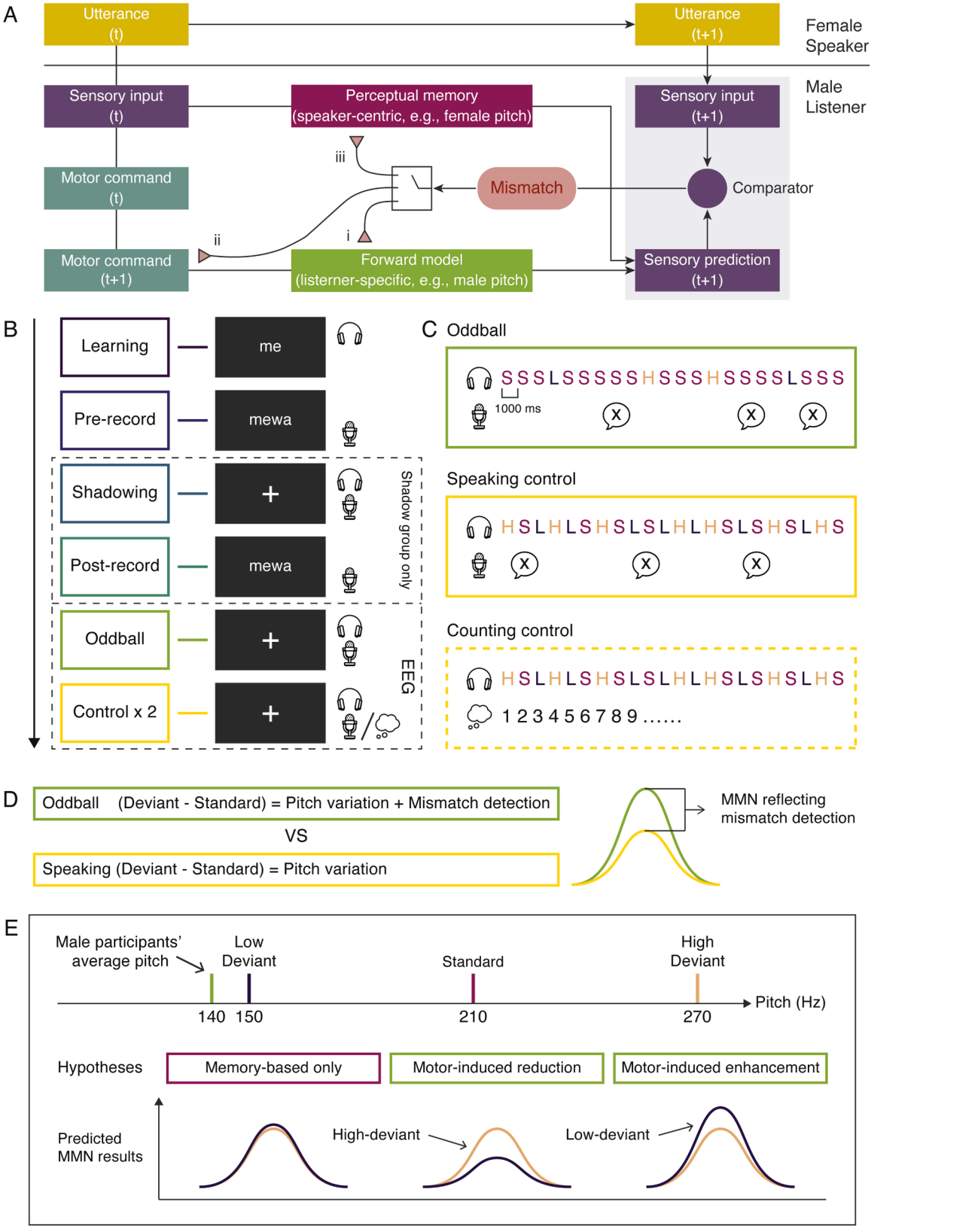
Beyond predicting our own speech, the integrated theory posits that we also predict others' speech using two types of predictions: memory- and motor-based, with the latter playing a central role in phonetic convergence (see Fig. 1A). Memory-based predictions are derived from external sources, storing information acquired through perceptual experiences such as hearing others speak or even non-speech events like the rustling of leaves. In the context of speech, these predictions are speaker-centric, capturing the unique vocal identity cues of individual speakers. In contrast, motor-based predictions originate from the listener's own speech production processes, utilizing forward models that govern self-generated vocal behaviour. Consequently, motor-based predictions are listener-specific, reflecting the listener's own vocal identity. Thus, memory- and motor-based predictions encapsulate different sub-phonemic information—one reflecting the external speaker's identity and the other reflecting the listener's identity. How do motor-based predictions contribute to phonetic convergence? The integrated theory proposes that, when hearing someone speak, listeners first engage in covert imitation, activating the motor commands needed to produce the same speech they hear, and then generate predictions about the speaker's upcoming utterances (see also Blakemore and Frith, 2005). Crucially, motor-based predictions not only capture the linguistic content of the speaker's speech but also reflect the listener's own vocal traits. Discrepancies between predicted and actual speech—resulting from individual differences in vocal tract characteristics and linguistic backgrounds—create prediction errors. These errors prompt adjustments in the listener's motor commands and forward models, driving their speech to converge with the speaker's vocal traits.

While the integrated theory offers a mechanistic account of phonetic convergence, evidence that supports the existence of motor-based predictions during speech perception or their role in facilitating phonetic convergence is rare. Eye-tracking studies suggest that listeners predict upcoming speech sounds uttered by others (Ito, 2024; Ito et al., 2018; X. Li et al., 2022; X. Li and Qu, 2024), and this predictive process is adaptable to specific speakers based on accent cues (Dahan et al., 2008; Johnson et al., 2022; Trude and Brown-Schmidt, 2012; Witteman et al., 2013). Ito et al. (2018), for example, had participants listen to highly predictable sentences (e.g., The tourists expected rain when the sun went behind the …) while viewing an array of objects that were phonologically related (clown) or unrelated (globe) to the target (cloud). Participants fixated on the sound-related object more than on the unrelated before hearing the target. More importantly, using accented speech to indicate speaker identity, Trude & Brown-Schmidt (Trude and Brown-Schmidt, 2012) observed predictive eye movements in participants before the complete utterances of the target words. These findings provide strong evidence that listeners predict upcoming speech sounds when hearing other speakers, and that these internal predictions seem to be speaker-centric and memory-base, rather than listener-specific motor-based, predictions.

Moreover, according to the integrated theory, motor-based predictions—whether when processing auditory feedback from one's own speech or input from other speakers—suppress auditory responses to signals that match the predictions. In other words, motor-based predictions reduce the brain's sensitivity to predicted acoustic features, thereby facilitating the detection of mismatches between predicted and actual input. However, recent research suggests that motor-based modulations of auditory processing are more multifaceted than previously thought. Studies have demonstrated that both imagined speaking and articulatory preparation enhance neural responses to matched speech sounds relative to mismatched ones, thereby enhancing sensitivity to predicted acoustic features (S. Li et al., 2020; Tian and Poeppel, 2013; Yang et al., 2024). These findings indicate that motor-based modulations on auditory processing exhibit varying functions depending on the specific articulatory activities engaged. Elucidating the functional roles of motor-based predictions in modulating auditory processing during speech perception is crucial for understanding how they contribute to phonetic convergence. Thus, clarifying whether and how motor-based predictions modulate auditory processing during speech perception, particularly in relation to phonetic convergence, is the key aim of the present study.

Auditory mismatch negativity (MMN) provides an effective methodological approach to investigate the functional role, specificity, and progression of internal predictions during phonetic convergence. Measured through electroencephalography (EEG), MMN is a negative deflection that occurs when the brain detects a mismatch between internal sound representations of frequently occurring 'standard' stimuli and unexpected 'deviant' inputs in an oddball paradigm (Garrido et al., 2009; Näätänen et al., 2005, 2007). Crucially, these internal sound representations are not mere reflections of the physical properties of the standard stimuli but are also shaped and modulated by listeners' experiences and attention, indicating top-down influences on auditory processing (Sussman et al., 2002; Tervaniemi et al., 2009). Previous MMN studies have demonstrated that listeners can form abstract voice representations despite constantly changing phonological information within a stimulus sequence (Di Dona et al., 2022; Tuninetti et al., 2017). These findings suggest that speaker-centric memory-based predictions contribute to the formation of internal sound representations and functionally reduce the brain's sensitivity to predicted acoustic features, thereby influencing how the brain processes and adapts to speech input.

In controlled laboratory settings, phonetic convergence is typically examined using a shadowing task, where participants quickly repeat each word they hear (Goldinger, 1998). However, these studies generally focus on behavioural changes rather than the underlying neural mechanisms. To address this gap, we implemented a novel speaking oddball task. In this task, participants listened to a series of standard stimuli interspersed with deviant stimuli and were randomly cued to repeat the standard. This design served two main purposes. First, the oddball component was employed to elicit stable internal representations of the standard stimuli, allowing us to use the deviant stimuli to probe the properties of these representations. Second, by incorporating the speaking element, the task mimicked the sensorimotor dynamic conditions of the traditional shadowing task, thereby eliciting similar top-down prediction signals that contribute to the formation of internal sound representations. This approach allowed us to investigate whether and how listener-specific motor-based predictions contribute to the



(caption on next page)

Fig. 1. Overview of theoretical model, experimental design, task configurations, and testing hypotheses. (A) Theoretical model of predictive processing in speech (adapted from the integrated theory with a focus on the sub-phonemic level which encodes individuals' vocal identity cues such as pitch). The speaker's utterance at time t provides sensory input to the listener, who covertly imitates the speaker's motor commands and generates motor-based predictions through a forward model of articulatory control. Simultaneously, memory-based predictions are derived from perceptual memory of the speaker's voice. At time $t + 1$, the speaker's new utterance is compared with these predictions; any mismatch leads to adjustments in the forward model, the motor commands, or the stored representation of the speaker's voice, facilitating dynamic adaptation and phonetic convergence. (B) Experimental procedure, including four behavioural tasks and three EEG tasks. The shadow group completed all tasks, while the non-shadow group did not participate in either shadowing or post-record task. (C) Illustrations of the three EEG tasks: the oddball task, the speaking control task, and the counting control task. In all tasks, S represents the standard stimulus, and H and L indicate the high and low deviant stimuli, respectively. In the oddball task, the standard stimuli were presented more often than the deviant stimuli that were interspersed randomly. In the speaking and counting control tasks, the three types of stimuli (originally as the standard, high deviant, and low deviant in the oddball task) were presented randomly with equal probability. During the oddball and speaking control tasks, participants listened to word stimuli and repeated them when prompted. In the counting control task, participants counted each stimulus covertly. In all tasks, stimuli were presented at a frequency of 1000 ms. (D) Conceptual diagram illustrating how the canonical MMN index was derived in the current study. The speaking control task was used as a baseline to counter out the factor of acoustic variations and derived the MMN responses that reflect mismatch detection. (E) The predicted results according to three distinct hypotheses. Above, illustration of the relations between the pitch levels of the stimuli and the average pitch of male participants. The standard reflects the speaker's vocal identity (female pitch) and hence yield the memory-based prediction in male listeners; whereas, the listener's vocal characteristics (male pitch) is the foundation of own production and hence the motor-based prediction. The relative distances either from the standard to the two deviants, or from male listener's pitch characteristics to the two deviants yield distinct predicted MMN results. Below, predicted MMN results based on three alternative hypotheses. If only memory-based predictions were generated, comparable MMN responses were expected for the low and high deviant stimuli. If motor-based predictions were generated and reduced sensitivity to listener-matched predicted acoustic features, a smaller MMN response was expected to the low deviant compared to the high deviant. Conversely, if motor-based predictions enhanced sensitivity to listener-matched predicted features, a larger MMN response was expected for the low deviant.

formation of internal sound representations of the standard stimuli and how they progress alongside convergence behaviour.

The current study aimed to determine whether and how motor-based predictions modulate auditory processing during speech perception and contribute to phonetic convergence. We examined both overall vocal convergence to novel word pronunciations and pitch convergence from male participants to female stimuli; the latter was critical in helping us explore the properties of internal sound representations in our EEG session. Disyllabic pseudo-Japanese words were employed to minimize lexical and other higher-level linguistic influences, and only male participants were recruited to respond to female stimuli in order to maximize the identity differences—here defined by gender. Participants were first presented with monosyllable sounds in a learning phase to familiarize themselves with the Japanese vowel sound inventory. Then, the shadowing task required participants to repeat the pseudo-words spoken by a female speaker, thereby providing them with complete exposure to her vocal identity cues such as pitch range. Participants' utterances were also recorded before and after the shadowing task (hence, pre- and post-record tasks), allowing us to assess whether vocal changes occurred and persisted independently of immediate auditory stimulation. Behavioural measures—including pitch and the initial 13 Mel-frequency cepstral coefficients (MFCCs)—were used to examine changes in vocal output following the shadowing task. MFCCs are parametric representations of acoustic signals obtained by applying a cosine transform to the logarithm of the short-term energy spectrum expressed on a Mel-frequency scale (Muda et al., 2010); they were used to indicate overall vocal convergence, while pitch indicate male participants' pitch convergence to female stimuli.

Then, we introduced a speaking oddball task that incorporated MMN measures to address our primary interest. MMN effects were used as a neurophysiological index for the mismatch between internal representations related to the standard stimuli and external deviant inputs. The speaking oddball task used a standard stimulus pitched at 210 Hz—approximating a female voice (Barkana and Zhou, 2015)—alongside low (150 Hz) and high (270 Hz) deviant stimuli, reflecting typical pitch divergence between male and female voices. By comparing MMN responses between the two pitch-modulated deviant stimuli, we sought to determine whether motor-based predictions reduce sensitivity to predicted acoustic features, thereby facilitating mismatch detection and vocal modification, or enhance sensitivity, potentially guiding perceptual learning and speech adaptation during phonetic convergence. Two EEG control tasks were also implemented to control for pitch-related perceptual differences inherent in our standard and deviant stimuli and to examine whether the observed effects in the main experiment persisted in the absence of matched covert vocalization (see Methods for details).

Following the integrated theory, we defined motor-based predictions as those reflecting the listener's own identity (i.e., male pitch in our study) and memory-based predictions as those reflecting the external speaker's identity (i.e., female pitch). Accordingly, we posited three alternative predictions (see Fig. 1). First, if listeners relied solely on memory-based predictions, the MMN responses to the low and high deviants would be comparable, as both reflected similar degrees of pitch deviation. Second, if motor-based predictions further suppressed responses to listener-matched pitch, the low deviant would elicit a smaller MMN than the high deviant. Third, conversely, if motor-based predictions enhanced responses to listener-matched pitch, the low deviant would elicit a larger MMN relative to the high deviant. We also conducted correlational analyses between our acoustic-behavioural measures (both MFCC and pitch) and neural responses (both ERP and MMN). If motor-based predictions occurred, we expected convergence behaviour to be correlated with neural responses related to the low, but not high, deviant. Lastly, participants who completed all tasks mentioned above comprised the shadow group. We further recruited a non-shadow control group that did not participate in the shadowing task. By comparing these groups, we aimed to determine how internal predictions, particularly motor-based predictions, evolve alongside phonetic convergence, offering insights into the adaptive mechanisms underpinning speech perception and production.

2. Methods

2.1. Participants

Sixty-two native Mandarin-speaking male participants were recruited. A final sample of 48 participants was included: 24 in the shadow group (ages 19–26, $M = 22.13$, $SD = 2.13$) and 24 in the non-shadow group (ages 19–28, $M = 21.96$, $SD = 2.39$). Four participants were excluded due to poor audio recordings, and ten were removed due to excessive noise in their EEG data. All participants reported normal hearing and no history of neurological or learning disorders. The study was approved by the institutional review board at New York University Shanghai and conducted in accordance with the Declaration of Helsinki. All participants provided informed consent and were compensated for their time.

2.2. Materials and stimuli

Stimuli for the shadowing task consisted of 40 disyllabic pseudo-Japanese words, which were also meaningless in Mandarin. These

words were selected from Yan et al. (2021) and recorded in a soundproof booth by a 22-year-old female native Mandarin speaker proficient in Japanese pronunciation. Recordings were made using Audacity software, a BETA58A Vocal Microphone, and a Tinsea mpa MINI preamplifier at a sampling rate of 44.1 kHz on a MacBook Pro. We processed the recordings in Praat (Boersma and Weenink, 2018) by first peak-normalizing each word recording to 99 % of its maximum amplitude and then scaling the sound pressure level to 70 dB SPL. Praat was also used to adjust the average pitch of each word to 210 Hz—representative of a typical female voice (Barkana and Zhou, 2015). The words (mean duration = 513 ms, SD = 6.7 ms) were divided into four sets of ten words each (list A, B, C, and D). One word was selected from each list to serve as standard stimuli (see Design and Procedure for more details). During EEG sessions, participants were randomly assigned with one of these standard words, which also underwent additional pitch adjustments to create low (150 Hz) and high (270 Hz) deviant stimuli.

In addition, 15 initial syllables were selected and extracted from the disyllabic word recordings to be used during the initial learning phase. These syllables had a mean duration of 351 ms (SD = 26.36 ms). They were pitch-shifted in 20 Hz increments from 130 Hz to 210 Hz, thereby spanning the typical male-to-female pitch range (Barkana and Zhou, 2015). Each syllable recording was peak-normalized to 99 % of its maximum amplitude and standardized to 70 dB SPL using Praat. The set included the Japanese vowels /a/, /i/, /u/, /e/, and /o/, with each vowel paired with three different consonants, yielding 15 unique syllables.

2.3. Design and procedure

The experiment included seven tasks (see Fig. 1): learning, pre-record, shadowing, post-record, speaking oddball, speaking control, and counting control. EEG signals were recorded during the last three tasks. Participants in the shadow group completed all seven tasks, while the non-shadow group was excluded from the shadowing and post-record tasks.

Participants began with the learning task, where they learned novel vowel pronunciations by listening to syllable recordings. In each trial, a syllable was displayed in white text at the centre of the screen. After pressing the ENTER key, participants heard the syllable five times at different pitch levels, randomized in order. They practiced silently and could replay the sequence as needed. In the pre-record task, participants read aloud disyllabic words displayed on the screen. Each word appeared in white for 600 ms before turning green, signalling the participant to speak. After uttering the word, participants pressed the SPACE bar to proceed. Each word was repeated three times, resulting in a total of 60 spoken responses (20 words, each repeated three times). Participants were instructed to mimic the vowel pronunciations they had learned in the learning task and to put the stress on the first syllable. During the shadowing task, participants repeated disyllabic words after hearing them. Each trial began with a white fixation, followed by auditory word playback 200 ms later. Once the word ended, the fixation turned green, cueing participants to repeat the word aloud. After repeating each word, they pressed the SPACE bar to continue. Each word was repeated three times. In total, participants completed 120 responses (20 words, each repeated six times across two blocks). The post-record task followed the same format as the pre-record, with participants reading aloud words presented on the screen. For each participant, three lists (e.g., list A, B, and C) were allocated to the pre-record, shadowing, and post-record tasks, respectively, while the fourth list (e.g., list D) was presented across all three tasks. This design ensured that convergence effects observed after the shadowing task would not be restricted to the directly shadowed words but would also generalize to new words that were not previously articulated or heard. List assignments were counterbalanced across participants.

In the speaking oddball task, participants listened to a sequence of repeated word stimuli while preparing to repeat the word when cued. Each stimulus was presented for approximately 350 ms, and stimuli were delivered with a stimulus onset asynchrony (SOA) of 1000 ms (1 Hz frequency). This SOA thus comprised the 350 ms stimulus duration followed by a 650 ms silent period before the next stimulus onset. A white fixation was displayed during the auditory stimuli, and it turned green to prompt participants to speak. Two pitch-varied deviant stimuli were randomly interspersed within the sequence of standard stimuli, with each deviant appearing after three to five standard stimuli. This resulted in a presentation probability of 80 % for standard stimuli and 10 % for each deviant. Each deviant occurred 60 times, and the standard stimulus occurred 480 times. Sixty responses were recorded in this task. Prior to each response, the standard stimuli were presented twice before the fixation turned green, leading to an additional 120 presentations of the standard stimuli. Different from the preceding behavioural tasks, only one word—and its two pitch-shifted variants—was used in the EEG sessions, and this word was selected from the word list employed in all previous tasks.

The speaking control task was similar to the oddball task but with standard and deviant stimuli presented with equal frequency (33. 3 % each). Each of the three stimuli (standard, low deviant, and high deviant) was presented 60 times in a pseudorandomized order, ensuring that no two consecutive sounds were identical. Participants responded vocally to 20 randomly selected standard stimuli after the fixation turned green. Thus, in this control task, any ERP differences among the three stimuli were attributed to pitch-related perceptual differences such as perceived intensity or naturalness. The counting control task used the same stimulus sequence as in the speaking control task and had participants covertly count the stimuli instead of vocalizing. The counting control task assessed whether pitch-related effects observed in the speaking control task persisted when there were linguistic content mismatches between covert speech motor activities and auditory inputs.

The experiment was written in MATLAB R2020b using the Psychophysics Toolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). Participants were seated in a soundproof, light-controlled booth, and auditory stimuli were delivered via ER-3C Insert Earphones connected to plastic air tubes.

2.4. Acoustic data processing and analysis

Vocal responses were recorded using a SHURE SM58 microphone and an MP13 Mini-Mic preamplifier, connected to a PC sound card at a sampling rate of 44.1 kHz and 16-bit resolution. Recording began at the onset of the green fixation. Utterance onsets and offsets were manually annotated using a customized MATLAB script. In the pre-record, shadowing, and post-record tasks, participants repeated each word three times. For analysis, the first utterance for each word was excluded due to frequent absences of sound outside the designated response window. In the shadowing task, only responses from the second half of the session were used for comparison with data collected from the pre-record task. This resulted in a final dataset where each participant from both groups contributed a same number of 40 responses for the pre-record task, 60 for the oddball task, and 20 for the speaking control task to obtain the same statistical power in comparisons between groups. Whereas the shadow group had 40 responses in each of the shadowing and post-record tasks to quantify the convergence effects.

Two dependent variables were used to assess vocal convergence: MFCC dissimilarity and pitch difference. The first 13 MFCCs were calculated for both participant and speaker utterances. Dynamic time warping was then applied to align the MFCC sequences, accounting for temporal differences between the utterances. Cosine distances between the aligned MFCCs were calculated, focusing on spectral similarity between the participant's and the speaker's utterances. The average cosine distance, termed MFCC dissimilarity, was used as a measure of vocal similarity, with lower values indicating greater similarity and thus more convergence. Pitch was extracted using a custom Praat script, ProsodyPro (Xu, 2013), and compared to the speaker's average pitch of 210

Hz. The difference between the speaker's and the participants' pitch, termed pitch difference, was calculated. A decrease in pitch difference after the shadowing task indicated convergence toward the female speaker's pitch. Given that the current study used disyllabic words with stress on the first syllable, we further examined pitch changes separately for the first and second syllables. This allowed us to determine whether any observed overall pitch change was primarily driven by male participants imitating the female pitch level or by an imitation of the speaker's stress pattern.

MFCC dissimilarity and pitch difference (for the whole utterance and separately for each syllable) were analysed using linear mixed-effects regression models (LMERs). All analyses were performed in R Statistical Software (R Core Team, 2022), with LMER analyses conducted using the lme4 package (version 1.1–31) (Bates et al., 2015). The fixed effect in the models was task (pre-record, shadowing, post-record, oddball, and speaking control for the shadow group; pre-record, oddball, and speaking control for the non-shadow group), with pre-record set as the reference level. We included both by-participant and by-word random effects, starting with a maximal random effects structure (Barr et al., 2013), which included random intercepts and random slopes for task when possible. We tested for singularity (i.e., near-zero variances or correlations indicating an overfitted or degenerate random effects structure) and convergence issues, simplifying the random effects structure until the model converged without singular fits. We followed a forward model-building strategy, testing whether the inclusion of task as a fixed effect improved model fit using chi-squared tests.

2.5. EEG data processing and analysis

EEG signals were recorded using a 32-channel active electrode system (Brain Vision actiCHamp, Brain Products) at a sampling rate of 1000 Hz in an electromagnetically shielded, soundproof booth. Electrodes were positioned according to the international 10–20 system using EasyCap, with impedance maintained below 10 kΩ. Data were recorded via Brain Vision PyCoder software and referenced to the Cz electrode. Horizontal and vertical electrooculograms (EOGs) were also recorded to monitor ocular activity. Custom scripts based on the FieldTrip toolbox (Oostenveld et al., 2011) were used for EEG data processing and analysis. The data were down sampled to 250 Hz and bandpass filtered with a Butterworth filter between 0.1 and 30 Hz. Channels with excessive noise or artifacts were repaired using the triangulation method to interpolate data from neighbouring channels. Independent component analysis (ICA) was then performed to identify and remove components related to eye blinks and movements, using the Extended Infomax algorithm from EEGLAB's runica method. Typically, 1–2 ICA components that matched EOG artifacts in timing, shape, and spatial distribution were removed. The data were then re-referenced to the average of all electrodes. Event-related potentials (ERPs) were computed for each trial over a 700-ms window, starting 100 ms before the stimulus onset. Each epoch was baseline-corrected to the 100-ms pre-stimulus interval. Trials with peak-to-peak amplitudes exceeding 100 μV in any 200-ms window were excluded, and additional trials with subtler artifacts such as muscle contamination were removed following visual inspection. To ensure a comparable number of trials across conditions, we only included standard trials that immediately preceded each deviant trial in the oddball task. All trials were retained in the two control tasks. Participants with >30 % of their trials rejected were excluded from further analysis. For the remaining participants, an average of 86 % of trials (corresponding to approximately 52 trials per stimulus type) were retained.

For raw MMN effects, auditory responses to low and high deviant stimuli were compared to the standards in the oddball task. In the speaking control task, we analysed ERP differences between deviant and standard stimuli to assess whether pitch-related acoustic variations elicited differential auditory responses. For the counting control task, ERP responses were compared between deviant and standard stimuli to determine if the pitch variation effects observed in the speaking control task persisted when covert speech motor activities did not align with the perceived words. Additionally, we performed two difference-in-difference (DID) analyses. The first DID analysis—calculated as oddball (deviant - standard) minus speaking (deviant - standard)—was employed to isolate the MMN effects that were specifically attributable to mismatch detection and were termed corrected MMN effects. This approach ensured that, if there was a difference between the MMN responses elicited by the two deviants, that difference stemmed from the discrepancy between each deviant and the standard rather than from perceptual disparities between the two deviants themselves (see Fig. 1C). The second DID analysis—speaking (deviant - standard) minus counting (deviant - standard)—assessed the interaction between covert speech motor activity and auditory linguistic content in response to pitch variations. Thus, our EEG analysis compared ERP responses to deviant versus standard stimuli within each task and further analysed deviant-standard ERP differences between tasks to break down the components of MMN effects observed in the current study.

For EEG analysis, we first conducted a temporal clustering permutation test (Maris and Oostenveld, 2007) focusing on the Fz channel, selected a priori as the representative site for MMN effects given its widespread recognition in the literature (Näätänen et al., 2007). We used a dependent samples *t*-test at each time point, identifying contiguous time segments in which the t-statistic exceeded a pre-cluster threshold of $p < .05$. For each such temporal cluster, we summed the t-values to form a cluster-level statistic, and then created a null distribution by randomly permuting condition labels 10,000 times. Any observed cluster whose summed t-value exceeded the 95th percentile of this null distribution was considered significant at the cluster level ($p < .05$). Next, we performed a spatiotemporal clustering permutation test across all electrodes to capture broader patterns of neural activity. In this analysis, temporal adjacency was defined by consecutive time points meeting the initial threshold ($p < .05$), while spatial adjacency required at least one neighbouring electrode. As in the single-channel test, a dependent samples *t*-test was used at each time-electrode point. We then summed the t-values for all contiguous points (in both time and space) and compared the resulting cluster-level statistics against a null distribution generated by 10,000 permutations. Clusters surpassing the 95th percentile of the null distribution were deemed significant at the cluster level ($p < .05$). Both temporal (Fz) and spatiotemporal (all electrodes) cluster-based tests were applied first to compare deviant versus standard ERP responses for each deviant type within each task. We then examined deviant–standard differences between tasks for each deviant type to explore the MMN effects and their interaction with covert speech motor activity and pitch-related acoustic variations. We also analysed our EEG data using a more conventional approach, extracting peak amplitudes within a defined time window and comparing these across conditions. See methods and results of this analysis in the Supplementary Material.

2.6. Acoustic-neural correlation

We also conducted a series of correlational analyses to examine the relationship between participants' vocal performance and neural responses in the oddball and speaking control tasks. We used two acoustic measures, MFCC dissimilarity and pitch difference, and two neurophysiological measures, ERP and MMN responses. Correlations were analysed between these two types of measures, focusing on ERP and MMN responses within the 70–230 ms window, where significant MMN effects were observed across all EEG tasks (see below).

3. Results

3.1. Acoustic data

For the shadow group, a total of 4800 vocal responses were initially collected. After excluding missed or incomplete responses, 4679

responses remained. We further removed trials exceeding three median absolute deviations (MADs) from each participant's median for each task. The MFCC dissimilarity analysis (4556 responses) revealed a significant task effect ($\chi^2(4) = 26.17$, $p < .001$). Greater similarities were observed in the shadowing ($\beta = -0.041$, $t = -5.55$, 95 % $CI = [-0.055, -0.026]$, $p < .001$) and post-record tasks ($\beta = -0.015$, $t = -2.08$, 95 % $CI = [-0.029, -0.001]$, $p = .037$) than in the pre-record task, indicating overall vocal convergence. The pitch difference analysis (4517 responses) also revealed a significant task effect ($\chi^2(4) = 17.90$, $p = .001$). The pitch difference was larger in the shadowing task than in the pre-record task ($\beta = 5.282$, $t = 2.52$, 95 % $CI = [1.19, 9.57]$, $p = .012$), suggesting pitch divergence. For the non-shadow group, 2880 vocal responses were collected, with 2799 responses retained after excluding missed or incomplete trials. No significant task effect was found for either MFCC dissimilarity ($\chi^2(2) = 0.29$, $p = .87$) or pitch difference ($\chi^2(2) = 2.51$, $p = .29$; see Fig. 2C-D).

To further explore the pitch divergence observed for the shadow group, we conducted separate pitch difference analyses for the first and second syllables. For the first syllable, adding task as a fixed effect improved the model fit ($\chi^2(4) = 18.60$, $p < .001$), but the following

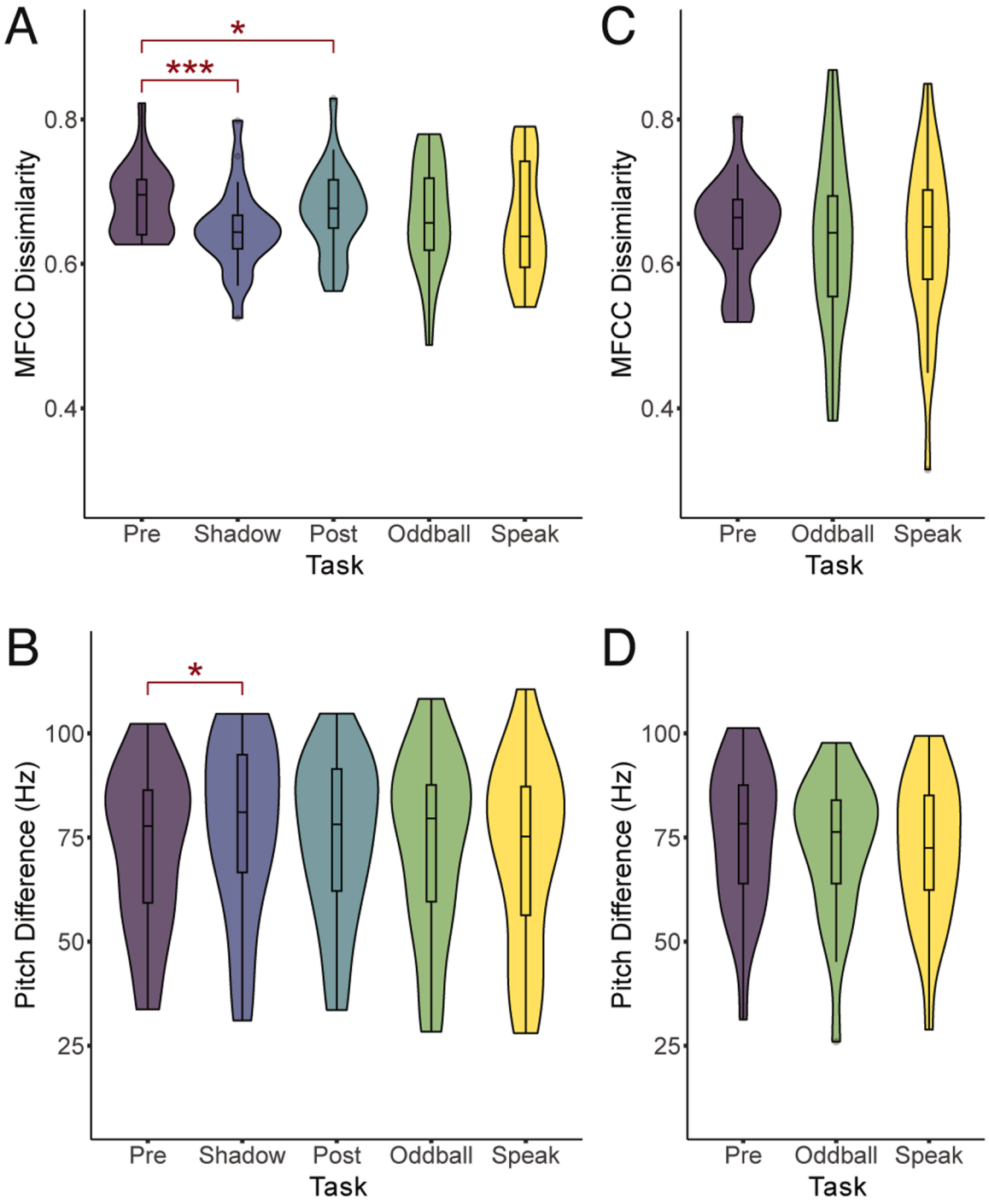


Fig. 2. Acoustic results for both shadow and non-shadow groups across various tasks. (A) MFCC dissimilarity and (B) pitch difference results for the shadow group. (C) MFCC dissimilarity and (D) pitch difference results for the non-shadow group. Each violin plot shows the distribution of data, with corresponding boxplots illustrating the median and interquartile range (IQR). Whiskers extend from the hinges to the farthest values within 1.5 times the IQR. The pre-record task serves as the reference baseline for comparisons, with decreases in values indicating convergence behaviour. ***$p < .001$, *$p < .05$.

LMER analysis only suggested a trend of increased pitch difference in the shadowing task compared to the pre-record task ($\beta = 4.428$, $t = 1.64$, 95 % $CI = [-0.86, 9.571$, $p = .101$). For the second syllable, adding task as a fixed effect did not improve the model fit ($\chi^2(4) = 5.97$, $p = .20$). Participants' first-second syllable pitch difference was 20.60 Hz ($CI = [14.81, 26.39]$) in the pre-record task, whereas the speaker's first-second syllable difference was 10.91 Hz ($CI = [6.77, 15.049]$). These results suggested that overall pitch divergence in the shadow group was likely due to participants imitating the speaker's stress pattern by lowering the first syllables' pitch.

3.2. EEG data

3.2.1. The shadow group

After confirming vocal convergence in the shadow group, we next examined whether motor-based predictions were generated during perception and, if so, whether these predictions suppressed or enhanced neural responses to listener-matched pitch stimuli (i.e., the low deviant). All results from the temporal and spatiotemporal cluster-based tests were reported following Sassenhagen & Draschkow (2019). Fig. 3A shows low deviant effects for the three EEG tasks. For the oddball task, the temporal cluster-based test revealed a significant difference at Fz between the low deviant and the standard from approximately 150 to 210 ms ($p = .020$, $t = -55.82$). The spatiotemporal test identified one positive cluster from approximately 110 to 230 ms over the left temporal and parietal regions ($p = .020$, $t = 4.83$) and one negative cluster from approximately 100 to 250 ms in the frontal and central areas ($p < .001$, $t = -5.23$). For the speaking control task, the temporal cluster-based test revealed a significant difference at Fz from approximately 90 to 170 ms ($p = .010$, $t = -61.25$). The spatiotemporal test revealed a negative cluster from approximately 90 to 170 ms in the frontal and central areas ($p = .006$, $t = -5.23$). For the counting control task, the temporal cluster-based test revealed a significant difference at Fz from approximately 130 to 210 ms ($p = .007$, $t = -75.23$). The spatiotemporal test identified a negative cluster from approximately 120 to 250 ms in the frontal and central regions ($p = .002$, $t = -5.37$). Our DID analysis showed that, when comparing the oddball and the speaking control tasks (see Fig. 3B), the spatiotemporal test revealed a negative cluster from approximately 120 to 210 ms across the frontal and central regions ($p = .011$, $t = -4.92$). Altogether, deviant-standard differences were observed for the low deviant across all three tasks, and this difference persisted even when contrasting the oddball and speaking control tasks. We interpreted the latter control-corrected difference as the genuine MMN response that reflected the detection of a mismatch between the internal standard sound representations and the external low deviant inputs.

Fig. 3C shows high deviant effects. For the oddball task, the temporal cluster-based test revealed no difference at Fz between the high deviant and the standard. However, the spatiotemporal test identified one positive cluster from approximately 490 to 600 ms over the parietal and central regions ($p = .005$, $t = 4.73$) and one negative cluster from approximately 470 to 600 ms in the right temporal and frontal regions ($p = .016$, $t = -4.94$). For the speaking control task, no difference was found at Fz, but the spatiotemporal test revealed a positive from approximately 170 to 280 ms involving central and midline channels ($p = .009$, $t = 4.09$). No significant difference was identified in the counting control task. Our DID analysis showed that, when comparing the two control tasks (see Fig. 3D), the spatiotemporal test revealed a positive cluster from approximately 180 to 300 ms over the central and parietal regions ($p = .008$, $t = 4.36$). Of our main interest, a deviant-standard difference was observed for the high deviant in the speaking control task, and this difference persisted when comparing the speaking control and counting control tasks. The latter finding suggested that the observed difference for the speaking control task only occurred when there was a content match between covert speech motor activity and received speech sounds. Our results from the shadow group seemed to support that motor-based predictions enhanced neural responses to listener-matched pitch stimuli.

3.2.2. The non-shadow group

The non-shadow group was included to explore how internal predictions modulated auditory processing when there was only limited exposure to the speaker's utterances. Comparing results between the shadow and the non-shadow groups allowed us to examine how internal predictions evolved with convergence behaviour. Fig. 4A shows low deviant effects for the non-shadow group. For the oddball task, the temporal cluster-based test revealed significant differences at Fz between the low deviant and the standard from approximately 70 to 230 ms ($p = .0026$, $t = -150.15$) and from approximately 370 to 600 ms ($p = .002$, $t = -178.21$). The spatiotemporal test revealed three positive and two negative clusters. Two positive clusters were localized to the occipital region from approximately 70 to 220 ms ($p = .011$, $t = 5.14$) and from approximately 500 to 600 ms ($p = .014$, $t = 5.13$). A third positive cluster involved channels in the left temporal and parietal areas from approximately 400 to 600 ms ($p = .007$, $t = 6.88$). Both negative clusters covered the frontal and central areas from approximately 70 to 260 ms ($p < .001$, $t = -5.90$) and from approximately 440 to 590 ($p = .003$, $t = -5.39$). For the speaking control task, the temporal cluster-based test revealed significant differences at Fz from approximately 80 to 200 ms ($p = .002$, $t = -92.20$) and from approximately 470 to 530 ms ($p = .015$, $t = -44.70$). The spatiotemporal test identified one positive and two negative clusters. The positive cluster involved the occipital region from approximately 430 to 560 ms ($p = .007$, $t = 6.11$). Both negative clusters spanned the frontal and central regions from approximately 70 to 220 ms ($p < .001$, $t = -5.63$) and from approximately 380 to 540 ms ($p < .001$, $t = -4.79$). For the counting control task, the spatiotemporal test revealed one negative cluster involving channels across the frontal, central, and left temporal regions from approximately 160 to 220 ms ($p = .024$, $t = -3.64$). No significant result was found for any DID analysis. Comparable deviant-standard differences were observed for the low deviant across all three tasks, meaning that the low deviant did not elicit the MMN response that we considered as reflecting the detection of a mismatch between the internal standard sound representations and the external low deviant inputs.

Fig. 4B shows high deviant effects. For the oddball task, the temporal cluster-based test revealed a significant difference at Fz between the high deviant and the standard from approximately 120 to 180 ms ($p = .009$, $t = -53.20$). The spatiotemporal test identified a negative cluster involving frontal and right temporal regions from approximately 120 to 190 ms ($p = .003$, $t = -4.53$). No significant cluster was found for either control task. Our DID analysis showed that, when comparing oddball and speaking control tasks (see Fig. 4C), the temporal cluster-based test identified a significant difference at Fz from approximately 120 to 190 ms ($p = .004$, $t = -53.01$). The spatiotemporal test also revealed a negative cluster involving the frontal, central, and right parietal regions from approximately 130 to 200 ms ($p = .004$, $t = -3.78$). A deviant-standard difference was observed for the high deviant in the oddball task and this difference persisted when contrasting the oddball and the speaking control tasks. In other words, the high deviant elicited the MMN response that we considered as reflecting the detection of a mismatch between the internal standard sound representations and the external high deviant inputs. The non-shadow group seemed to show different results from the shadow group, and we discuss these findings below.

3.3. Acoustic-neural correlation

Correlational analyses were performed to examine whether the degree of vocal convergence was related to neural responses. MFCC dissimilarity was positively correlated with ERP peak amplitude at Fz elicited by the low deviant in the oddball task ($r = 0.59$, $p = .002$; see Fig. 5A) and in the speaking control task ($r = 0.41$, $p = .049$; see Fig. 5B),

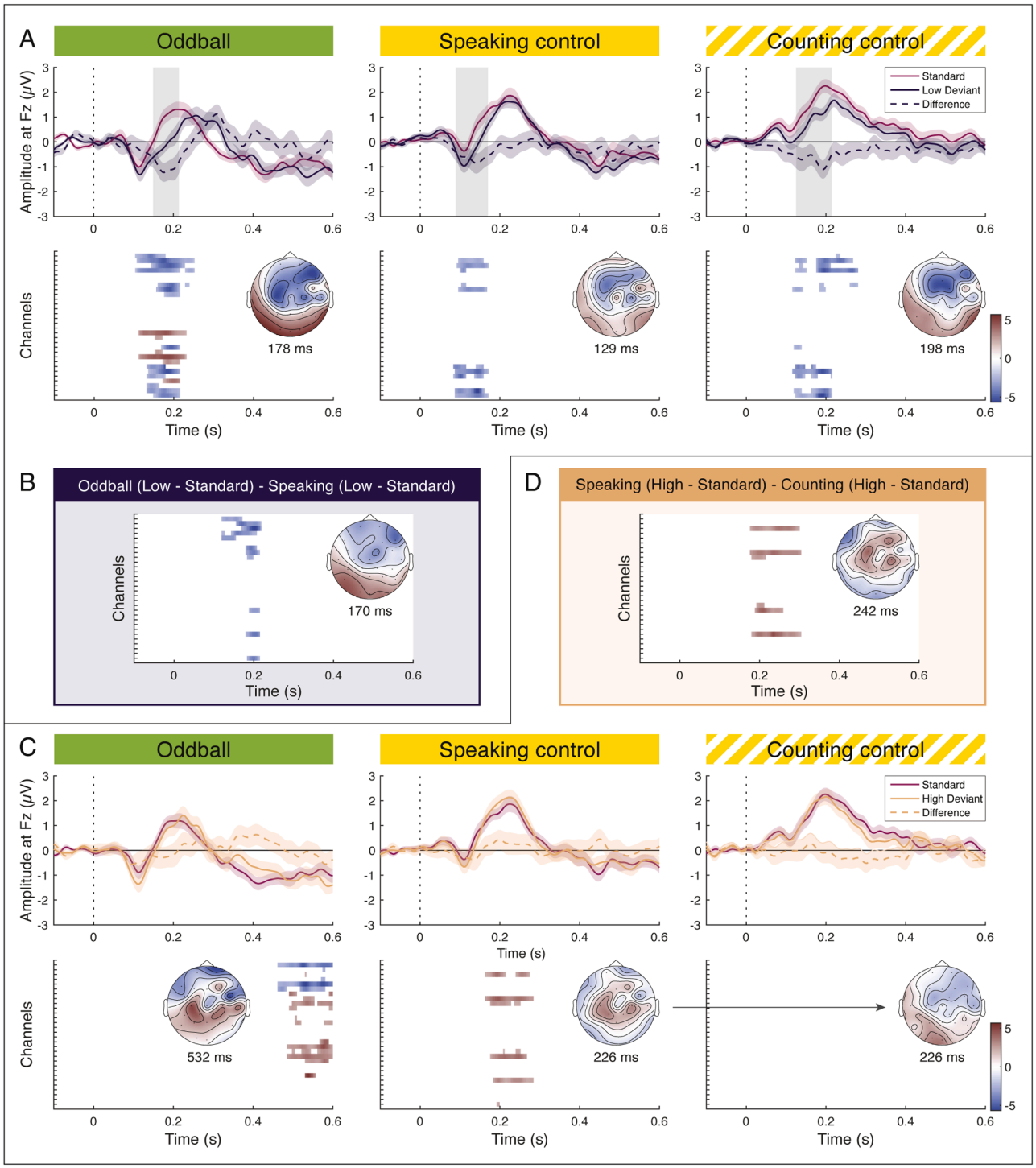


Fig. 3. ERP results for the shadow group. (A) Low deviant effects. Top Panel: ERP responses at Fz, showing responses to low deviant and standard stimuli for the oddball, speaking control, and counting control tasks (solid lines). Dotted lines represent ERP differences, with shaded areas indicating significant periods ($p < .05$, cluster-level). Bottom Panel: Spatiotemporal characteristics of ERP differences between low deviant and standard stimuli across all channels, with significant clusters indicated by blue (negative amplitude) and red (positive amplitude) bars ($p < .05$, cluster-corrected). Topographic displays at the centroid time points are shown in the upper right. (B) Spatiotemporal characteristics of ERP differences calculated by subtracting the speaking control task responses (low deviant - standard) from the oddball task, with a topographic display at the specified time point. (C) High deviant effects. Similar to panel A, showing ERP responses to high deviant stimuli at Fz, with corresponding ERP differences. Bottom Panel: Spatiotemporal characteristics of ERP differences between high deviant and standard stimuli, with topographic displays at centroid time points. (D) Spatiotemporal characteristics of ERP differences calculated by subtracting counting control task responses (high deviant - standard) from those in the speaking control task, with a topographic display at the centroid time point.

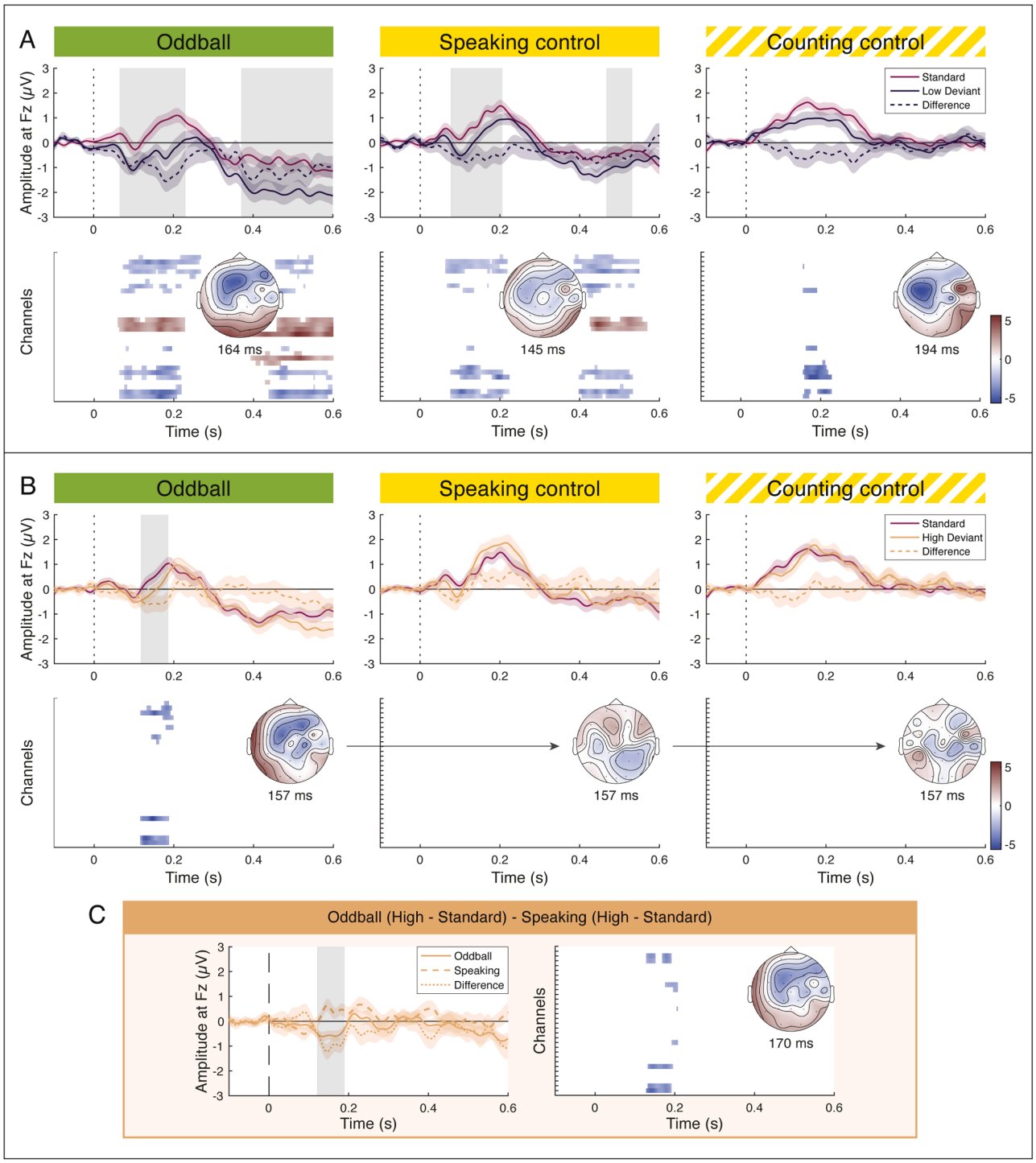


Fig. 4. ERP results for the non-shadow group. (A) Low deviant effects. Top Panel: ERP responses at Fz, showing responses to low deviant and standard stimuli for the oddball, speaking control, and counting control tasks (solid lines). Dotted lines represent ERP differences, with shaded areas indicating significant periods ($p <$ .05, cluster-level). Bottom Panel: Spatiotemporal characteristics of ERP differences between low deviant and standard stimuli across all channels, with significant clusters indicated by blue (negative amplitude) and red (positive amplitude) bars ($p <$ .05, cluster-corrected). Topographic displays at the centroid time points are shown in the upper right. (B) High deviant effects. Similar to panel A, showing ERP responses to high deviant stimuli at Fz, with corresponding ERP differences. Bottom Panel: Spatiotemporal characteristics of ERP differences between high deviant and standard stimuli across all channels, with topographic displays at centroid time points. (C) ERP difference responses at Fz and spatiotemporal characteristics of ERP differences calculated by subtracting the speaking control task responses (high deviant - standard) from the oddball task. Solid and dashed lines represent ERP difference responses for the oddball and speaking control tasks, respectively, with the dotted line representing the difference of differences between the two tasks. A topographic display of significant clusters at the centroid time point is included.

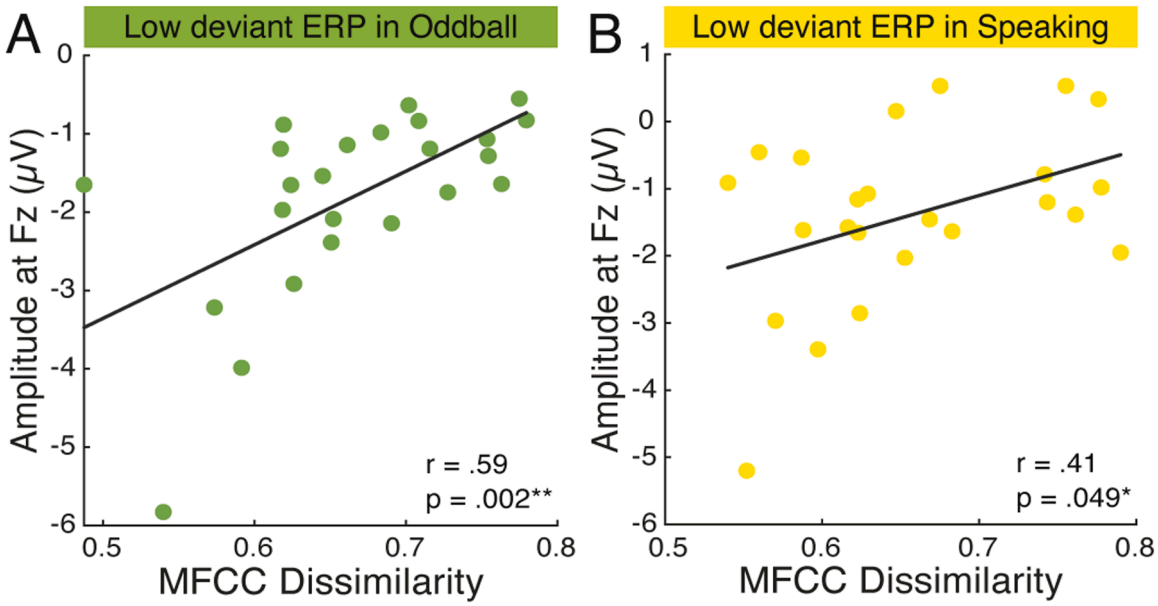


Fig. 5. Correlational results. (A) Correlation between MFCC dissimilarity and ERP peak amplitude at Fz elicited by the low deviant in the oddball task for the shadow group. (B) Correlation between MFCC dissimilarity and ERP peak amplitude at Fz elicited by the low deviant in the speaking control task for the shadow group. Each scatterplot includes fitted trend lines, with corresponding correlation coefficients (r) and p-values displayed.

indicating that participants whose vocalizations more closely resembled the speaker's exhibited larger negative ERP responses to the low deviant. This result also supported that motor-based predictions enhanced the brain's sensitivity to predicted acoustic features, with greater enhancement associated with closer resemblance of participants' vocalizations to the speaker's. No significant correlation was observed for the non-shadow group.

4. Discussion

Phonetic convergence is thought to result from adaptive corrections to prediction errors that arise from discrepancies between internal motor-based predictions and external speech inputs. However, the precise modulatory effects of these predictions on auditory processing and their contributions to convergence behaviour remain unclear. In this study, we aimed to determine whether motor-based predictions reduce the brain's sensitivity to predicted acoustic features—thus facilitating mismatch detection and vocal modification—or instead enhance such sensitivity to support perceptual learning, which may subsequently guide vocal adjustment during speaking. We used MMN effects, recorded during the speaking oddball task and corrected using the speaking control task, as a neurophysiological index of the mismatch between internal standard sound representations and external deviant inputs, thereby providing insights into internally generated auditory predictions.

Our results revealed significant vocal convergence in the shadow group, alongside pitch divergence during the shadowing task—a pattern likely driven by participants' imitation of the speaker's stress pattern. EEG analyses for the shadow group showed significant raw and corrected MMN effects for the low deviant but no significant MMN effect for the high deviant, consistent with the hypothesis that motor-based predictions enhance responses to listener-matched pitch. In other words, listener-specific motor-based predictions appear to enhance the brain's sensitivity to expected acoustic features. Moreover, our correlational analyses indicated that this enhancement was stronger in participants whose vocalizations more closely approximated the speaker's, suggesting a link between convergence strength and neural sensitivity. Conversely, the non-shadow group exhibited a significant raw, but not corrected, MMN effect for the low deviant and both significant raw and corrected MMN effects for the high deviant. We interpreted these findings as evidence that, with limited exposure to the speaker's voice, the non-shadow group's neural responses were primarily driven by acoustic factors rather than by internal motor-based or memory-based predictions. Overall, our findings suggested that motor-based predictions enhance the brain's sensitivity to predicted acoustic features, thereby contributing to perceptual learning that guides vocal modification during speech.

4.1. Vocal imitation and pitch divergence

Our behavioural results revealed effective vocal learning of novel pronunciations in the shadow group, which extended to new words not encountered during the shadowing task. Previous research on phonetic convergence in non-native speech shows that learners can imitate specific phonetic features absent in their native languages, such as voice onset time, consonant glottalization, vowel duration, and formant frequencies (Flege and Eefting, 1988; Jiang and Kennison, 2022; Llompart and Reinisch, 2019; Podlipský and Simácková, 2015; Rojczyk et al., 2023; Zając and Rojczyk, 2014). However, the persistence and generalizability of these imitation effects seem to be less pronounced than those observed in our study. This difference might stem from previous studies' focus on specific acoustic attributes that accentuate differences between languages. In contrast, our use of MFCC measurements to assess overall vocal performance captured a broader array of acoustic measures, likely contributing to more robust imitation effects. In studies involving native sounds, convergence behaviours vary depending on the acoustic features which participants focus on, and changes in individual attributes do not always align with overall vocal performance (Babel and Bulatov, 2012; Pardo et al., 2017). By using feature-inclusive MFCC measurements, we therefore observed more robust vocal learning effects, as participants likely imitated different acoustic features to varying degrees. The use of a single word in the EEG sessions could not capture the full spectrum of vocal imitation, hence explaining the nonsignificant vocal change in either the oddball or speaking control task.

In the shadow group, contrary to our expectations, we observed pitch divergence. Further exploration revealed that participants tended to lower their pitch on the stressed syllables, aligning their stress patterns

more closely with the speaker's. Pitch serves both linguistic (intonation, lexical tones) and paralinguistic functions (age, gender, emotion). Studies on pitch imitation have reported mixed results, often showing small and variable effects (Babel and Bulatov, 2012; Garnier et al., 2013; Gentilucci and Bernardis, 2007; Kappes et al., 2009; Lewandowski and Nygaard, 2018; Pardo et al., 2013, 2017; Sato et al., 2013). Aubanel and Nguyen (2020) observed that robust pitch imitation typically results from systematic manipulation of pitch variations within model speakers' utterances (Garnier et al., 2013; Kappes et al., 2009; Sato et al., 2013). Our findings of stress pattern imitation aligned with this tendency, suggesting that phonetic convergence is more influenced by adaptable articulatory patterns than by fixed traits like gender. Although unspecific group effects due to inherent speaker differences across groups cannot be entirely ruled out, our focus on within-group pitch changes in the shadow group demonstrates clear vocal plasticity. This adaptability supports phonetic convergence's social function in fostering cohesion and regulating identity within interactions (Dragojevic et al., 2015; Giles et al., 1991), underscoring the flexibility of vocal communication in adapting to social contexts.

4.2. Functional specificity and progression of internal predictions during phonetic convergence

In this study, motor- and memory-based predictions of others' speech were defined as distinct signals reflecting the listener's and the speaker's vocal identities, respectively. We proposed that these two types of predictions modulate auditory processing in distinct ways, each becoming more refined as phonetic convergence occurs, though in different manners. Motor-based predictions, rooted in precise one-to-one mappings between sensory and motor systems, are hypothesized to enhance the brain's sensitivity to acoustic features that match the listener's own vocal characteristics. As convergence progresses, these signals enable more precise detection of subtle phonetic variations, yielding stronger neural responses to listener-matched (predicted) features relative to unpredicted ones. In contrast, memory-based predictions, acquired through exposure to the speaker's voice, tend to reduce the brain's sensitivity to features that align with the speaker's vocal identity. This suppression of neural responses serves to diminish redundant processing when the input conforms to expected speaker characteristics. Moreover, as convergence progresses, memory-based predictions become less precise but more adaptable, allowing the system to accommodate natural fluctuations in the speaker's voice rather than demanding exact replication. Below, we first discuss the modulatory effects attributed to memory-based predictions, followed by those of motor-based predictions. Given that memory-based predictions are an automatic and inescapable component of speech perception (Kukona, 2020; Pickering and Gambi, 2018), it is essential to account for their influence when interpreting motor-based effects.

4.3. Memory-based predictions: reducing sensitivity to predicted features and adapting to speaker variability

Memory-based predictions in this study likely consisted of two components: short-term acoustic memories encoding the standard stimuli's acoustic properties and long-term perceptual memories related to gender categorization. The short-term memories helped detect both low and high deviants but were subject to change through shadowing experience, while the long-term memories facilitated detection of the low deviant which always signalled a gender shift from the preceding stimulus in all EEG tasks. As a result, deviant-standard ERP differences were larger for the low deviant than for the high deviant across all EEG tasks in both groups. Studies using passive listening oddball tasks have shown that listeners are sensitive to pitch changes even in the presence of variations in formant frequencies which are critical for vowel categorization (Di Dona et al., 2022; Tuninetti et al., 2017). These findings supported our results by suggesting that, with or without exposure to a speaker's voice, listeners can rely on long-term memories to automatically detect gender-related pitch variations.

More importantly, the raw MMN effect was not significant for the high deviant in the shadow group, suggesting that shadowing experience led to reduced sensitivity to this deviant which fell within the same gender category as the standard. Given the variability of human voices, listeners must generalize across these variations to form a stable percept of identity (Lavan, Burston, et al., 2019, 2019). Research has shown that training with non-native vowels enhances MMN, while training with new voices reduces it, indicating that the processes retrieving linguistic and identity information are influenced differently by experience (Di Dona et al., 2021). In our study, memory-based predictions following the shadowing task likely became more tolerant of the speaker's acoustic variations. This tolerance reduced sensitivity to the high deviant, which might have been perceived as a different word produced by the same speaker. The high deviant was generated by shifting the pitch up while keeping all other formant information identical to the standard. Since formant information is crucial for reflecting the speaker's vocal tract characteristics, the high deviant might have been interpreted as the same speaker producing the word with a different pitch contour, thus potentially a different word. The positive cluster observed for the high deviant in the speaking control task likely reflected this process. Previous oddball studies have shown that active listening enhances P3 responses compared to passive listening (Bekinschtein et al., 2009; Di Dona et al., 2021; Justen and Herbert, 2018; Rutiku et al., 2024). Our speaking control task could be interpreted as a pseudo-oddball task, where the high deviant was perceived as a lexical deviant, triggering P3-like responses due to its relevance to the word repetition task.

In contrast, the raw MMN effect was significant for the high deviant in the non-shadow group, suggesting preserved sensitivity to this stimulus which was acoustically different from the standard. With limited exposure to the speaker's voice, memory-based predictions—particularly those derived from short-term acoustic memories—likely encoded the standard stimuli's acoustic properties more faithfully, thus preserving sensitivity to the high deviant. The late negativity component observed for the low deviant in the non-shadow group might reflect late discriminative negativity (LDN), a component more pronounced in children than adults and more prominent for speech than nonspeech signals (Bishop et al., 2011; David et al., 2020; Di Dona et al., 2022; Liu et al., 2014). LDN is associated with the formation of sound regularities in memory, particularly for meaningful phonological information, which is why it is more pronounced in children still developing their linguistic abilities. LDN responses in the non-shadow group might indicate an ongoing learning process as participants formed representations of the speaker's voice and gender-related sound patterns during the task.

4.4. Motor-based predictions: enhancing sensitivity to predicted features and reflecting vocal resemblance to the speaker

In contrast to memory-based predictions, which reduce sensitivity to predicted, speaker-matched acoustic features, our results suggested that motor-based predictions enhance sensitivity to predicted, listener-matched features. In the shadow group, both raw and corrected MMN effects were observed for the low deviant, suggesting that the shadowing experience enhanced sensitivity to pitch levels within the typical male range. Previous studies have shown that imagined speaking and articulatory preparation enhance neural responses to expected speech sounds compared to unexpected ones (S. Li et al., 2020; Tian and Poeppel, 2013), and this enhancement extends to low-level acoustic features like loudness, even when the imagined and received speech do not align in linguistic content (Tian et al., 2018). In line with these findings, the deviant-standard differences observed for the low deviant in both speaking and counting control tasks, along with the lack of differences between the two control tasks, suggested that motor-based modulations can enhance sensitivity to specific acoustic features, such as pitch,

regardless of whether the internal articulatory activity matches the received speech content.

Furthermore, in the shadow group, the significant corrected MMN effect for the low deviant—obtained by subtracting the speaking control responses from those in the oddball task—suggested that this effect specifically reflected mismatch detection between internal sound representations and external deviant inputs, independent of physical acoustic properties or perceptual factors such as loudness or naturalness. In other words, while the raw MMN effect might be influenced by these confounding factors, the corrected MMN more accurately captured the contribution of motor-based predictive processes. For speech perception, motor-based efference copies are proposed to function as top-down attentional modulators, increasing the gain of responses to expected sensory features and sharpening tuning selectivity (Hickok et al., 2011). In our oddball task, the repeated presentation of the standard stimulus likely reinforced participants' internal predictions, resulting in more precise motor-based predictions. Our correlation analyses further revealed that participants whose vocalizations more closely approximated the speaker's exhibited stronger responses to the low deviant, and this relationship was more pronounced in the oddball task than in the speaking control task. These findings suggested that more effective vocal learning is associated with the formation of more accurate motor-based predictions, thereby enhancing sensitivity to listener-matched pitch features.

In the non-shadow group, the absence of significant differences across the three EEG tasks suggested that either no motor-based predictions were generated or, if they were, they were imprecise. In line with the results from the shadow group—where motor-based predictions enhanced responses to the low pitch—the low-standard ERP differences observed in these EEG tasks could also reflect motor-based enhancements. However, it was equally possible that these effects were driven by memory-based predictions related to gender categorization. While memory-based predictions might have played a role in enhancing responses to the low deviant, they could not fully account for the corrected MMN effect observed for the low deviant in the shadow group. Taken together, our results from the two groups supported that motor-based predictions enhance sensitivity to predicted, listener-matched acoustic features and become more precise as convergence behaviour occurs.

Overall, our EEG results indicated that memory- and motor-based predictions contribute distinctly to auditory processing during speech perception and to phonetic convergence. Our findings suggested that extensive exposure to a speaker's voice is critical for refining internal predictions: memory-based predictions that are derived from perceptual exposure tend to suppress neural responses to features that match the speaker's vocal identity, whereas motor-based predictions that are acquired through articulatory experiences enhance sensitivity to features aligned with the listener's own vocal characteristics. This nuanced perspective contrasts with earlier views that motor-based predictions primarily reduce sensitivity to expected features. Instead, our results indicate that motor-based predictions potentially facilitate perceptual learning during sustained auditory exposure, particularly for detailed acoustic properties, while immediate vocal repetition further fine-tunes motor commands based on these learned targets, thereby promoting phonetic convergence. Insights from music research support this: listening to and producing music activates auditory-limbic pathways, with music training focusing on detailed sound perception in a rewarding context (Blood et al., 1999; Kraus and White-Schwoch, 2015; Salimpoor et al., 2013). Given the social nature of phonetic convergence, enhanced sensitivity to a speaker's voice may be linked to cognitive and reward networks, suggesting that motor-based predictions not only modulate auditory processing but also engage reward-based learning to foster convergence.

4.5. Methodological validation and limitations

We employed a speaking oddball paradigm which diverged from traditional passive listening tasks where participants' attention is typically directed away from stimuli. Our interactive setup, where participants took turns articulating with the same speaker, was designed to engage sensorimotor processes and foster predictive mechanisms. In addition, unlike the typical active oddball tasks where participants had to respond to deviants (Bennington and Polich, 1999; Justen and Herbert, 2018; Sussman et al., 1998, 2002), we instructed participants to repeat the word when the sound stopped, making the deviants task-irrelevant. As a result, we did not observe the P3 components typically associated with the stimulus-driven orienting of attention to novel stimuli (Nieuwenhuis et al., 2011; Polich, 2007), except the high deviant effects observed in the speaking control task in the shadow group (see discussion above). Our task design allowed participants to generate consistent predictions about the speaker's utterances without actively anticipating deviants.

To address concerns about whether deviant-standard differences observed in the speaking oddball task reflected a genuine mismatch detection process or were confounded by physical differences between stimuli and neural refractoriness (Jacobsen and Schröger, 2001, 2003; Näätänen et al., 2007; Schröger and Wolff, 1996), we controlled for physical differences between the two deviants by comparing auditory responses between the oddball and speaking control tasks. We employed a DID approach, as attention and/or motor engagement likely differed across EEG tasks. Compared to the oddball task, vocal responses were more frequent in the speaking control task, and participants covertly counted each stimulus in the counting control task. Alongside this, N1 responses appeared to be reduced across these tasks. Corollary discharge, a type of signals generated through motor-to-sensory transformations, has been shown to suppress auditory responses during action and action preparation, regardless of whether auditory inputs match expected sensory outcomes (Horváth et al., 2012; S. Li et al., 2020; Sanmiguel et al., 2013; Schneider et al., 2014). We interpreted the N1 reduction across EEG tasks as the result of one type of predictive motor activity that uniformly dampened auditory responses, regardless of linguistic alignment between motor signals and external auditory inputs. Crucially, we acknowledged that the DID approach, while helpful for controlling confounds, may not completely isolate the neural signatures of prediction error from other task-related influences. The differences in participants' mental states across tasks might introduce additional variability, which we attempted to mitigate by focusing on deviant-standard differences rather than on raw ERP amplitudes. Further research is needed to replicate these findings and to develop more refined experimental designs that can better illustrate the interplay between various internal prediction signals and external inputs.

In addition to our primary cluster-based permutation analyses, we also performed a conventional ANOVA on peak amplitude differences computed within a fixed 20-ms window centred on the grand-average negative peak (see Supplementary Material). Although the overall pattern was consistent with our permutation results, the ANOVA yielded non-significant differences between conditions. We believed this could be due to several factors. First, the ANOVA approach—relying on a fixed time window—might not adequately capture the dynamic and temporally distributed nature of prediction error signals that were revealed by our continuous, cluster-based analysis. Second, our unconventional speaking oddball task, in which participants were randomly cued to vocalize, likely engaged multiple top-down processes (e.g., motor-based and memory-based predictions) that modulated auditory processing in complex and time-varying ways. Averaging neural activity within a narrow window could dilute these dynamic effects, leading to non-significant differences in the conventional analysis. Although our methods deviated from the conventional MMN literature, the combination of the speaking oddball task and cluster-based permutation tests provided a novel approach to reveal the multifaceted and dynamic

interplay between various internal prediction signals and external inputs. Future research may consider refining the task design and, more importantly, employing more alternative methods to fully characterize these effects, thereby providing further validation of the current theoretical model of phonetic convergence.

The present study implemented a novel paradigm to investigate the fundamentals of phonetic interaction. How these processes operate in relatively more dynamic, naturalistic conditions of everyday conversation calls for further investigation. In natural dialog, speakers engage in continuous turn-taking and mutual adaptation, as demonstrated by Mukherjee et al. (2019) who showed that phonetic convergence is modulated by distinct alpha and beta oscillatory dynamics in the speaker and listener. Our study employed controlled tasks, such as the shadowing and speaking oddball paradigms, to demonstrate the basic operations, yet the full spectrum of neural processes occurring in real-world interactions remains to be explored. Moreover, the oscillatory responses in specific frequency bands may illustrate more neural dynamics as demonstrated in Mukherjee et al. (2019). Future work would be informative to consider neural responses in the spectral domain in combination with naturalistic experimental designs to provide a more comprehensive account of the sensorimotor and predictive mechanisms underlying phonetic convergence in ecologically valid settings.

5. Conclusion

In conclusion, this study highlights the distinct roles that memory- and motor-based predictions play in modulating auditory processing during speech perception and their contributions to phonetic convergence. Specifically, our findings indicate that memory-based predictions suppress the brain's sensitivity to predicted acoustic features related to the speaker's vocal traits, while motor-based predictions enhance sensitivity to features aligned with the listener's vocal traits. The study critically delineates the modulatory roles of motor-based predictions in auditory processing. This increased sensitivity to predicted features contrasts with the traditional view that motor-based predictions primarily reduce sensitivity to predicted features. Furthermore, the study suggests that motor-based and memory-based predictions undergo different refinement processes during phonetic convergence: motor-based predictions become more precise over time, while memory-based predictions become less precise but more adaptable, allowing for greater flexibility in accommodating the natural variations in the speaker's voice. Together, these refinements facilitate perceptual learning and enable more accurate vocal adjustments. Overall, our results emphasize the importance of both types of predictions, as well as their interplay, in understanding how individuals adapt their speech to align with others, with broader implications for theories of speech perception, communication, and sensorimotor learning.

Funding

This study was supported by the National Natural Science Foundation of China 32271101, Program of AI-Driven Initiative to Promote Research Paradigm Reform and Empower Disciplinary Advancement by Shanghai Municipal Education Commission (SMEC), Program of Introducing Talents of Discipline to Universities, Base B16018, and NYU Shanghai Boost Fund.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the readability and language of the manuscript.

CRediT authorship contribution statement

Yuchunzi Wu: Writing – original draft, Visualization, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhili Han:** Writing – review & editing, Methodology, Conceptualization. **Xing Tian:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to express our gratitude to Zihan Zhou, Shyla Zhou, and Yichen Guan for their assistance with data collection. Additionally, we thank Yuqi Su, Yi Yao, and Aidan Huang for their valuable contributions to acoustic data processing. Their efforts were instrumental in the successful completion of this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2025.121169.

Data availability

Data will be shared upon request.

References

Adank, P., Hagoort, P., Bekkering, H., 2010. Imitation improves language comprehension. Psychol. Sci. 21 (12), 1903–1909. https://doi.org/10.1177/0956797610389192.

Aubanel, V., Nguyen, N., 2020. Speaking to a common tune: between-speaker convergence in voice fundamental frequency in a joint speech production task. PLoS. One 15 (5), e0232209. https://doi.org/10.1371/journal.pone.0232209.

Babel, M., Bulatov, D., 2012. The role of fundamental frequency in phonetic accommodation. Lang. Speech. 55 (2), 231–248. https://doi.org/10.1177/0023830911417695.

Barkana, B.D., Zhou, J., 2015. A new pitch-range based feature set for a speaker's age and gender classification. Applied Acoustics 98, 52–61. https://doi.org/10.1016/j.apacoust.2015.04.013.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1). https://doi.org/10.18637/jss.v067.i01.

Behroozmand, R., Larson, C.R., 2011. Error-dependent modulation of speech-induced auditory suppression for pitch-shifted voice feedback. BMC. Neurosci. 12 (1), 54. https://doi.org/10.1186/1471-2202-12-54.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L., 2009. Neural signature of the conscious processing of auditory regularities. Proc. Natl. Acad. Sci. 106 (5), 1672–1677. https://doi.org/10.1073/pnas.0809667106.

Bennington, J.Y., Polich, J., 1999. Comparison of P300 from passive and active tasks for auditory and visual stimuli. International Journal of Psychophysiology 34 (2), 171–177. https://doi.org/10.1016/S0167-8760(99)00070-7.

Bishop, D.V.M., Hardiman, M.J., Barry, J.G., 2011. Is auditory discrimination mature by middle childhood? A study using time-frequency analysis of mismatch responses from 7 years to adulthood. Dev. Sci. 14 (2), 402–416. https://doi.org/10.1111/j.1467-7687.2010.00990.x.

Blakemore, S.-J., Frith, C., 2005. The role of motor contagion in the prediction of action. Neuropsychologia 43 (2), 260–267. https://doi.org/10.1016/j.neuropsychologia.2004.11.012.

Blood, A.J., Zatorre, R.J., Bermudez, P., Evans, A.C., 1999. Emotional responses to pleasant and unpleasant music correlate with activity in paralimbic brain regions. Nat. Neurosci. 2 (4), 382–387. https://doi.org/10.1038/7299.

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer* (Version 6.0.40). http://www.praat.org/.

Bosshardt, H.-G., Sappok, C., Knipschild, M., Hölscher, C., 1997. Spontaneous imitation of fundamental frequency and speech rate by nonstutterers and stutterers. J. Psycholinguist. Res. 26 (4), 425–448. https://doi.org/10.1023/A:1025030120016.

Brainard, D.H., 1997. The Psychophysics Toolbox. *Spat Vis* 10, 433–436.

Brouwer, S., Mitterer, H., Huettig, F., 2010. Shadowing reduced speech and alignment. J. Acoust. Soc. Am. 128 (1), EL32–EL37. https://doi.org/10.1121/1.3448022.
Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., Hari, R., 2000. Speaking modifies voice-evoked activity in the human auditory cortex. Hum. Brain Mapp. 9 (4), 183–191. https://doi.org/10.1002/(SICI)1097-0193(200004)9:4<183::AID-HBM1>3.0.CO;2-Z.
Dahan, D., Drucker, S.J., Scarborough, R.A., 2008. Talker adaptation in speech perception: adjusting the signal or the representations? Cognition 108 (3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003.
David, C., Roux, S., Bonnet-Brilhault, F., Ferré, S., Gomot, M., 2020. Brain responses to change in phonological structures of varying complexity in children and adults. Psychophysiology. 57 (9). https://doi.org/10.1111/psyp.13621.
Delvaux, V., Soquet, A., 2007. The influence of ambient speech on adult speech productions through unintentional imitation. Phonetica 64 (2–3), 145–173. https://doi.org/10.1159/000107914.
Di Dona, G., Scaltritti, M., Sulpizio, S., 2021. Early differentiation of memory retrieval processes for newly learned voices and phonemes as indexed by the MMN. Brain Lang. 220, 104981. https://doi.org/10.1016/j.bandl.2021.104981.
Di Dona, G., Scaltritti, M., Sulpizio, S., 2022. Formant-invariant voice and pitch representations are pre-attentively formed from constantly varying speech and non-speech stimuli. European Journal of Neuroscience 56 (3), 4086–4106. https://doi.org/10.1111/ejn.15730.
Dragojevic, M., Gasiorek, J., Giles, H., 2015. Communication accommodation theory. The International Encyclopedia of Interpersonal Communication. Wiley, pp. 1–21. https://doi.org/10.1002/9781118540190.wbeic006.
Flege, J.E., Eefting, W., 1988. Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation. J. Acoust. Soc. Am. 83 (2), 729–740. https://doi.org/10.1121/1.396115.
Gambi, C., Pickering, M.J., 2013. Prediction and imitation in speech. Front. Psychol. 4 (June), 340. https://doi.org/10.3389/fpsyg.2013.00340.
Garnier, M., Lamalle, L., Sato, M., 2013. Neural correlates of phonetic convergence and speech imitation. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00600.
Garrido, M.I., Kilner, J.M., Stephan, K.E., Friston, K.J., 2009. The mismatch negativity: a review of underlying mechanisms. Clinical Neurophysiology 120 (3), 453–463. https://doi.org/10.1016/j.clinph.2008.11.029.
Gentilucci, M., Bernardis, P., 2007. Imitation during phoneme production. Neuropsychologia 45 (3), 608–615. https://doi.org/10.1016/j.neuropsychologia.2006.04.004.
Giles, H., Coupland, N., Coupland, J., 1991. Accommodation theory: communication, context, and consequence. Contexts of Accommodation. Cambridge University Press, pp. 1–68. https://doi.org/10.1017/CBO9780511663673.001.
Goldinger, S.D., 1998. Echoes of echoes? An episodic theory of lexical access. Psychol. Rev. 105 (2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251.
Greenlee, J.D.W., Jackson, A.W., Chen, F., Larson, C.R., Oya, H., Kawasaki, H., Chen, H., Howard, M.A., 2011. Human auditory cortical activation during self-vocalization. PLoS. One 6 (3), e14744. https://doi.org/10.1371/journal.pone.0014744.
Heinks-Maldonado, T.H., Mathalon, D.H., Houde, J.F., Gray, M., Faustman, W.O., Ford, J.M., 2007. Relationship of imprecise corollary discharge in schizophrenia to auditory hallucinations. Arch. Gen. Psychiatry 64 (3), 286. https://doi.org/10.1001/archpsyc.64.3.286.
Hickok, G., 2012. The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. J. Commun. Disord. 45 (6), 393–402. https://doi.org/10.1016/j.jcomdis.2012.06.004.
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.
Horváth, J., Maess, B., Baess, P., Tóth, A., 2012. Action–Sound coincidences suppress evoked responses of the Human auditory cortex in EEG and MEG. J. Cogn. Neurosci. 24 (9), 1919–1931. https://doi.org/10.1162/jocn_a_00215.
Houde, J.F., Nagarajan, S.S., Sekihara, K., Merzenich, M.M., 2002. Modulation of the auditory cortex during speech: an MEG study. J. Cogn. Neurosci. 14 (8), 1125–1138. https://doi.org/10.1162/089892902760807140.
Ito, A., 2024. Phonological prediction during comprehension: a review and meta-analysis of visual-world eye-tracking studies. J. Mem. Lang. 139, 104553. https://doi.org/10.1016/j.jml.2024.104553.
Ito, A., Pickering, M.J., Corley, M., 2018. Investigating the time-course of phonological prediction in native and non-native speakers of English: a visual world eye-tracking study. J. Mem. Lang. 98, 1–11. https://doi.org/10.1016/j.jml.2017.09.002.
Jacobsen, T., Schröger, E., 2001. Is there pre-attentive memory-based comparison of pitch? Psychophysiology. 38 (4), 723–727. https://doi.org/10.1111/1469-8986.3840723.
Jacobsen, T., Schröger, E., 2003. Measuring duration mismatch negativity. Clinical Neurophysiology 114 (6), 1133–1143. https://doi.org/10.1016/S1388-2457(03)00043-9.
Jiang, F., Kennison, S., 2022. The impact of L2 English learners' Belief about an interlocutor's English proficiency on L2 phonetic accommodation. J. Psycholinguist. Res. 51 (1), 217–234. https://doi.org/10.1007/s10936-021-09835-7.
Johnson, E.K., van Heugten, M., Buckler, H., 2022. Navigating accent variation: a developmental perspective. Annu Rev. Linguist. 8 (1), 365–387. https://doi.org/10.1146/annurev-linguistics-032521-053717.
Justen, C., Herbert, C., 2018. The spatio-temporal dynamics of deviance and target detection in the passive and active auditory oddball paradigm: a sLORETA study. BMC. Neurosci. 19 (1), 25. https://doi.org/10.1186/s12868-018-0422-3.
Kappes, J., Baumgaertner, A., Peschke, C., Ziegler, W., 2009. Unintended imitation in nonword repetition. Brain Lang. 111 (3), 140–151. https://doi.org/10.1016/j.bandl.2009.08.008.
Kleiner, M., Brainard, D., Pelli, D., 2007. "What's New in Psychtoolbox-3?" Perception 36 ECVP Abstract Supplement.
Kraus, N., White-Schwoch, T., 2015. Unraveling the biology of auditory learning: a cognitive–Sensorimotor–Reward framework. Trends Cogn. Sci. (Regul. Ed.) 19 (11), 642–654. https://doi.org/10.1016/j.tics.2015.08.017.
Kukona, A., 2020. Lexical constraints on the prediction of form: insights from the visual world paradigm. Journal of Experimental Psychology: Learning, Memory, and Cognition 46 (11), 2153–2162. https://doi.org/10.1037/xlm0000935.
Lavan, N., Burston, L.F.K., Garrido, L., 2019a. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. British Journal of Psychology 110 (3), 576–593. https://doi.org/10.1111/bjop.12348.
Lavan, N., Burton, A.M., Scott, S.K., McGettigan, C., 2019b. Flexible voices: identity perception from variable vocal signals. Psychon. Bull. Rev. 26 (1), 90–102. https://doi.org/10.3758/s13423-018-1497-7.
Lewandowski, E.M., Nygaard, L.C., 2018. Vocal alignment to native and non-native speakers of English. J. Acoust. Soc. Am. 144 (2), 620–633. https://doi.org/10.1121/1.5038567.
Li, S., Zhu, H., Tian, X., 2020. Corollary discharge versus efference copy: distinct neural signals in speech preparation differentially modulate auditory responses. Cerebral Cortex. https://doi.org/10.1093/cercor/bhaa154.
Li, X., Li, X., Qu, Q., 2022. Predicting phonology in language comprehension: evidence from the visual world eye-tracking task in Mandarin Chinese. Journal of Experimental Psychology: Human Perception and Performance 48 (5), 531–547. https://doi.org/10.1037/xhp0000999.
Li, X., Qu, Q., 2024. Verbal working memory capacity modulates semantic and phonological prediction in spoken comprehension. Psychon. Bull. Rev. 31 (1), 249–258. https://doi.org/10.3758/s13423-023-02348-5.
Liu, H.-M., Chen, Y., Tsao, F.-M., 2014. Developmental changes in mismatch responses to Mandarin consonants and lexical tones from early to middle childhood. PLoS. One 9 (4), e95587. https://doi.org/10.1371/journal.pone.0095587.
Llompart, M., Reinisch, E., 2019. Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. Lang. Speech. 62 (3), 594–622. https://doi.org/10.1177/0023830918803978.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164 (1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.
Muda, L., Begam, M., Elamvazuthi, I., 2010. Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques.
Mukherjee, S., Badino, L., Hilt, P.M., Tomassini, A., Inuggi, A., Fadiga, L., Nguyen, N., D'Ausilio, A., 2019. The neural oscillatory markers of phonetic convergence during verbal interaction. Hum. Brain Mapp. 40 (1), 187–201. https://doi.org/10.1002/hbm.24364.
Näätänen, R., Jacobsen, T., Winkler, I., 2005. Memory-based or afferent processes in mismatch negativity (MMN): a review of the evidence. Psychophysiology. 42 (1), 25–32. https://doi.org/10.1111/j.1469-8986.2005.00256.x.
Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: a review. Clinical Neurophysiology 118 (12), 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026.
Nieuwenhuis, S., De Geus, E.J., Aston-Jones, G., 2011. The anatomical and functional relationship between the P3 and autonomic components of the orienting response. Psychophysiology. 48 (2), 162–175. https://doi.org/10.1111/j.1469-8986.2010.01057.x.
Niziolek, C.A., Nagarajan, S.S., Houde, J.F., 2013. What does motor efference copy represent? Evidence from speech production. Journal of Neuroscience 33 (41), 16110–16116. https://doi.org/10.1523/JNEUROSCI.2137-13.2013.
Olmstead, A.J., Viswanathan, N., Aivar, M.P., Manuel, S., 2013. Comparison of native and non-native phone imitation by English and spanish speakers. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00475.
Olmstead, A.J., Viswanathan, N., Cowan, T., Yang, K., 2021. Phonetic adaptation in interlocutors with mismatched language backgrounds: a case for a phonetic synergy account. J. Phon. 87, 101054. https://doi.org/10.1016/j.wocn.2021.101054.
Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). *FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data* (pp. 1–9). https://doi.org/10.1155/2011/156869.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382–2393. https://doi.org/10.1121/1.2178720.
Pardo, J.S., Jordan, K., Mallari, R., Scanlon, C., Lewandowski, E., 2013. Phonetic convergence in shadowed speech: the relation between acoustic and perceptual measures. J. Mem. Lang. 69 (3), 183–195. https://doi.org/10.1016/j.jml.2013.06.002.
Pardo, J.S., Urmanche, A., Wilman, S., Wiener, J., 2017. Phonetic convergence across multiple measures and model talkers. Attention, Perception, & Psychophysics 79 (2), 637–659. https://doi.org/10.3758/s13414-016-1226-0.
Pelli, D.G., 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis. 10, 437–442.
Pickering, M.J., Gambi, C., 2018. Predicting while comprehending language: a theory and review. Psychol. Bull. 144 (10), 1002–1044. https://doi.org/10.1037/bul0000158.
Pickering, M.J., Garrod, S., 2013. An integrated theory of language production and comprehension. Behavioral and Brain Sciences 36 (04), 329–347. https://doi.org/10.1017/S0140525x12001495.
Podlipský, V.J., & Simácková, S. (2015). Phonetic imitation is not conditioned by preservation of phonological contrast but by perceptual salience. *ICPhS*.

Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clinical Neurophysiology 118 (10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019.

Rojczyk, A., Sturm, P., Przedlacka, J., 2023. Phonetic imitation in L2 speech: immediate imitation of English consonant glottalization by speakers of Polish. Lang. Acquis. 1–12. https://doi.org/10.1080/10489223.2023.2253545.

Rutiku, R., Fiscone, C., Massimini, M., Sarasso, S., 2024. Assessing mismatch negativity (MMN) and P3b within-individual sensitivity—A comparison between the local–global paradigm and two specialized oddball sequences. European Journal of Neuroscience 59 (5), 842–859. https://doi.org/10.1111/ejn.16302.

Salimpoor, V.N., van den Bosch, I., Kovacevic, N., McIntosh, A.R., Dagher, A., Zatorre, R. J., 2013. Interactions between the nucleus accumbens and auditory cortices predict music reward value. Science (1979) 340 (6129), 216–219. https://doi.org/10.1126/science.1231059.

Sanmiguel, I., Todd, J., Schröger, E., 2013. Sensory suppression effects to self-initiated sounds reflect the attenuation of the unspecific N1 component of the auditory ERP. Psychophysiology. 50 (4), 334–343. https://doi.org/10.1111/psyp.12024.

Sassenhagen, J., Draschkow, D., 2019. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology. 56 (6). https://doi.org/10.1111/psyp.13335.

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J.-L., Nguyen, N., 2013. Converging toward a common speech code: imitative and perceptuo-motor recalibration processes in speech production. Front. Psychol. 4. https://doi.org/10.3389/fpsyg.2013.00422.

Schneider, D.M., Nelson, A., Mooney, R., 2014. A synaptic and circuit basis for corollary discharge in the auditory cortex. Nature 513 (7517), 189–194. https://doi.org/10.1038/nature13724.

Schröger, E., Wolff, C., 1996. Mismatch response of the human brain to changes in sound location. Neuroreport 7 (18), 3005–3008. https://doi.org/10.1097/00001756-199611250-00041.

Sussman, E., Ritter, W., Vaughan, H.G., 1998. Attention affects the organization of auditory input associated with the mismatch negativity system. Brain Res. 789 (1), 130–138. https://doi.org/10.1016/S0006-8993(97)01443-1.

Sussman, E., Winkler, I., Huotilainen, M., Ritter, W., Näätänen, R., 2002. Top-down effects can modify the initially stimulus-driven auditory organization. Cognitive Brain Research 13 (3), 393–405. https://doi.org/10.1016/S0926-6410(01)00131-8.

Team, R. C., 2022. R: A language and Environment For Statistical Computing. R Foundation for Statistical Computing.

Tervaniemi, M., Kruck, S., De Baene, W., Schröger, E., Alter, K., Friederici, A.D., 2009. Top-down modulation of auditory processing: effects of sound context, musical expertise and attentional focus. European Journal of Neuroscience 30 (8), 1636–1642. https://doi.org/10.1111/j.1460-9568.2009.06955.x.

Tian, X., Ding, N., Teng, X., Bai, F., Poeppel, D., 2018. Imagined speech influences perceived loudness of sound. Nat. Hum. Behav. 2 (3), 225–234. https://doi.org/10.1038/s41562-018-0305-8.

Tian, X., Poeppel, D., 2013. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. J. Cogn. Neurosci. 25 (7), 1020–1036. https://doi.org/10.1162/jocn_a_00381.

Trude, A.M., Brown-Schmidt, S., 2012. Talker-specific perceptual adaptation during online speech perception. Lang. Cogn. Process. 27 (7–8), 979–1001. https://doi.org/10.1080/01690965.2011.597153.

Tuninetti, A., Chládková, K., Peter, V., Schiller, N.O., Escudero, P., 2017. When speaker identity is unavoidable: neural processing of speaker identity cues in natural speech. Brain Lang. 174, 42–49. https://doi.org/10.1016/j.bandl.2017.07.001.

Ventura, M.I., Nagarajan, S.S., Houde, J.F., 2009. Speech target modulates speaking induced suppression in auditory cortex. BMC. Neurosci. 10 (1), 58. https://doi.org/10.1186/1471-2202-10-58.

Wagner, M.A., Broersma, M., McQueen, J.M., Dhaene, S., Lemhöfer, K., 2021. Phonetic convergence to non-native speech: acoustic and perceptual evidence. J. Phon. 88, 101076. https://doi.org/10.1016/j.wocn.2021.101076.

Whitford, T.J., 2019. Speaking-induced suppression of the auditory cortex in humans and its relevance to schizophrenia. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 4 (9), 791–804. https://doi.org/10.1016/j.bpsc.2019.05.011.

Witteman, M.J., Weber, A., McQueen, J.M., 2013. Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. Attention, Perception, & Psychophysics 75 (3), 537–556. https://doi.org/10.3758/s13414-012-0404-y.

Xu, Y., A Tool for Large-scale Systematic Prosody Analysis, 2013. In: In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), pp. 7–10.

Yan, Y., Yang, Y., Ando, M., Liu, X., Kambara, T., 2021. Multisensory connections of novel linguistic stimuli in Japanese as a native language and referential tastes. Eur. J. Investig. Health Psychol. Educ. 11 (3), 999–1010. https://doi.org/10.3390/ejihpe11030074.

Yang, F., Zhu, H., Cao, X., Li, H., Fang, X., Yu, L., Li, S., Wu, Z., Li, C., Zhang, C., Tian, X., 2024. Impaired motor-to-sensory transformation mediates auditory hallucinations. PLoS. Biol. 22 (10), e3002836. https://doi.org/10.1371/journal.pbio.3002836.

Zając, M., Rojczyk, A., 2014. Imitation of English vowel duration upon exposure to native and non-native speech. Poznan Stud. Contemp. Linguistics 50 (4). https://doi.org/10.1515/psicl-2014-0025.