Regulation of sensorimotor serial learning in vocal pitch production for speech by motor compensation rather than sensory error

Yuhan Lu[1,2,6], Xiaowei Tang[3,6], Zhenyan Xiao[2,4], Anqi Xu[5], Junxi Chen[3]**, Xing Tian[1,2,4,7]*

[1]Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, China

[2]NYU-ECNU Institute of Brain and Cognitive Science, New York University Shanghai, Shanghai 200062, China

[3]Department of Neurosurgery, Guangzhou Sanjiu Brain Hospital, Guanzhou 510510, China

[4]Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning; Division of Arts and Sciences, New York University Shanghai, Shanghai 200122, China

[5]School of Humanities and Social Sciences, Harbin Institute of Technology, Shenzhen 518055, China

[6]These authors have contributed equally to this work

[7]Lead Contact

Corresponding:
*Xing Tian, E-mail: xing.tian@nyu.edu

**Junxi Chen, E-mail: chenjunxi2006@163.com

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Xing Tian (xing.tian@nyu.edu).

Abstract

Motor control is essential for organisms to efficiently interact with the environment by maintaining accurate action and adjusting to future changes. Speech production, one of the most complex motor behaviors, relies on a feedback control process to detect sensory errors and trigger updates in a feedforward control process that implements compensations. However, the specific contributions of these critical processes in sensorimotor learning during continuous vocal production remain debated. Here, we used two experimental designs in five experiments to dissociate these mechanisms. First, we employed a serial-dependence design with randomized pitch perturbations, dissociating the influences of sensory errors and motor compensation on subsequent vocalizations on a trial-by-trial basis. We found that motor compensation, rather than sensory errors, predicted the compensatory responses in the subsequent trials, suggesting instantaneous serial learning mediated by updates in the feedforward process. This compensation-driven serial learning was generalized across productions of different vowel categories. Second, we further implemented the experimental design in a sentential context, where auditory perturbation occurred only on the preceding syllable. Any learning effects in the subsequent syllable without pitch perturbation would reflect changes in the feedforward process. Our results consistently revealed that compensation in the preceding syllable predicted pitch changes in the subsequent syllable, but only when the two adjacent syllables were embedded within a word boundary. Collectively, the study provides ecological-valid evidence supporting that error-based motor compensation, incorporating cognitive and linguistic constraints, directly regulates the feedforward process and mediates the instantaneous serial learning in successive actions.

Keywords: speech motor control, feedforward, feedback, internal forward model, inverse model, sensory error, motor compensation, sensorimotor learning, serial dependence

Introduction

The accuracy and precision of motor output, as a result of efficient interplay among motor, sensory and cognitive systems, are crucial for human survival and social interactions.[1-3] In the domain of speaking, for example, the control of articulators relies on real-time integration among the language system, speech motor system and sensory feedback from the external environment.[4-7] Specifically, the brain anticipates the sensory consequences of articulatory movements via internal forward models,[8-13] continuously compares the predictions to sensory re-afference in the feedback process (**Fig. 1A**).[14-16] Perturbations in sensory feedback, whether occurring naturally as in tongue twisters or introduced experimentally, elicit sensory errors that trigger the inverse model and update motor programming in the feedforward control to correct articulation in real-time (**Fig. 1AB**).[17,18] Notably, such online compensatory adjustments can propagate to influence subsequent productions,[19-21] reflecting a form of rapid, trial-by-trial sensorimotor learning.

It remains heavily debated, however, how these serial adjustments contribute to sensorimotor learning via feedback and feedforward control mechanisms.[5,22-25] One possibility is that serial adjustments reflect changes in the feedback control process, whereby sensory errors encountered in a previous utterance recalibrate the sensory system's sensitivity to subsequent errors.[26,27] Supporting this hypothesis, recent findings show that exposure to unreliable auditory feedback down-weights auditory error signals and leads to reduced compensatory response in subsequent production.[19,28] Alternatively, another possibility posits that serial learning reflects updates to the feedforward process, and motor compensation acts not only as an immediate corrective response, but also leaves short-term traces in speech sound map or articulator map (as proposed in the DIVA model).[22,25,29,30] These traces can persist beyond the removal of perturbation, leading to relatively long-lasting adaptive changes that shape subsequent productions even in the absence of feedback perturbation and sensory errors.[20]

Yet, disentangling the contribution of feedback and feedforward processes to sensorimotor learning remains a major challenge, as sensory errors and motor compensations are tightly coupled. Standard auditory feedback perturbation and adaptation paradigms cannot easily differentiate the two because of their consistent co-occurrence across trials.[31] To overcome these obstacles, we first employed a novel serial-dependence paradigm with randomized pitch perturbations to isolate the trial-by-trial influences of sensory errors and motor compensations on subsequent speech productions.[32,33] Next, we modified and extended the experimental design into a sentential context, where we selectively applied auditory perturbation to a preceding syllable, remaining its subsequent syllable unperturbed. This design, by removing the auditory perturbation in the current production and hence the feedback process, directly investigated whether updates in the feedforward process mediate the instantaneous serial learning. Our results in five experiments consistently show that prior motor compensation, rather than sensory error, predicts subsequent vocal production. The serial learning effects generalized across different vowel categories but were constrained by sentential structures. These results provide evidence indicating that motor compensation is not only a reflexive response to sensory perturbation, but also drives updates in feedforward control, forming an instantaneous plasticity under the linguistic constraints during continuous speaking.

Results

Experiment 1: Trial-to-trial learning in repeated production of identical vowel

In the first experiment, we used the simplest setting where participants repeatedly produced a single vowel continuously for 2.5 sec, while their auditory feedback was unexpectedly downward or upward perturbed (*N* = 14; **Fig. 1B**). Through a long sequence of trials in which six levels of perturbation amount were randomly applied to each trial (**Fig. 1C**), we can separately explore how the serial changes of the current trial depended on the sensory error and motor compensation of the immediately preceding trial. Throughout the study, we defined each within-trial perturbation and compensation as C_t and P_t, and its preceding-1-trial perturbation and compensation as C_{t-1} and P_{t-1}.

We first measured motor responses to perturbation in the current trial (i.e., the relationship between C_t and P_t). We averaged across trials with the same amount of perturbation. It demonstrated that motor compensations were opposed to the direction of pitch perturbation, started around 150 ms after the onset of the perturbation in the conditions of -300 cents, -180 cents, -60 cents, and 60 cents pitch shift (**Fig. 1D**, from left to right, respectively, *ts*(13) = 36.13, 56.36, 109.85, and -92.86, *ps* = 0.043, 0.021, 0.008, and 0.024, cluster-based permutation test), yet this effect was not significant in the condition of 180-cent and 300-cent perturbations (*ps* > 0.203). To compare conditions, we selected 150~250 ms window that was commonly used to quantify the amount of motor compensation.[21,34] Compensation amounts did not significantly differ across upward and downward perturbations ($p > 0.067$, Friedman's chi-square test, FDR corrected; **Fig. 1E**).

<Figure 1 is about here>

Next, we tested the motor-compensation and sensory-error hypotheses by separately measuring the dependency of preceding-1-trial sensory error or motor compensation on its subsequent trial. Specifically, we characterized if C_t (in each trial) would follow or oppose C_{t-1} or P_{t-1} (in its preceding trial) (**Fig. 2A**). Therefore, we plotted C_t as a function of C_{t-1} and P_{t-1} - P_t, respectively, in both trial level and participant level (**Fig. 2BC**, left). The averaged data across trials and participants demonstrated that both functions yielded a derivative-of-Gaussian-shaped (DoG) curve (ps = 0.0001 for both, two-sided bootstrap, FDR corrected) (**Fig. 2BC,** right). That is, C_t in the current trial followed either C_{t-1} or P_{t-1} - P_t in the preceding trial, especially around 0; if C_{t-1} or P_{t-1} - P_t were approaching the extreme condition (far from 0), the following effect became weak. It suggested that within-trial compensation was likely to be updated from both preceding sensory error and motor compensation. The serial learning was quantified as the height of the DoG curve (shown as grey double arrows).

Caution should be taken, however, that the serial effect of P_{t-1} - P_t could be merely driven by the opposing effect of P_t on C_t, as shown in **Fig. 1DE.** Therefore, we further combined trials with the same P_t to isolate the dependency of P_{t-1} on C_t. The main effect of P_{t-1} was not significant ($\chi^2(5) < 6.65$, $p > 0.247$; Friedman's chi-square test, FDR corrected), suggesting that the preceding sensory error only contributed to the instant computation for compensation, but not directly to the future production. We further built a linear mixed-effect model to assess the contribution of P_t, P_{t-1}, and C_{t-1} to the C_t. The results revealed a significantly negative coefficient of P_t and a significantly positive coefficient of

C_{t-1} (t(11,680) = -21.8 and 3.03, p = 0.0001 and 0.004, respectively, one-sample t-test, FDR corrected; **Fig. 2E**). No effect was found for P_{t-1} (t(11,680) = 1.55, p = 0.160). These results suggest that C_{t-1}, but not P_{t-1}, had a serial effect on C_t.

<Figure 2 is about here>

Experiment 2: Trial-to-trial learning across the production of different vowels

Speech production involved variable vowel categories and thus in Experiment 2, we explored whether the serial effect of C_{t-1} remained across different vowels. In Experiment 2a, participants (N = 17) were asked to produce four different vowels that were randomly presented in each trial, i.e., 啊 (/a1/), 咿 (/i1/), 呃 (/e1/), and 哦 (/o1/). We observed a similar serial effect of C_{t-1} on C_t (p = 0.0001, two-sided bootstrap, FDR corrected; **Fig. 3A**), but not for the effect of P_{t-1} (**Fig. S1**). When comparing experiments 2a and 1, the amplitude of the DoG curve was significantly lower when participants produced different vowels than when they produced the same vowel (p = 0.036, two-sided bootstrap, FDR corrected; **Fig. 3C**). The reduction of serial learning can be caused by different vowels or by the uncertainty of next vowel. To exclude the latter possibility, we conducted Experiment 2b where the four vowels were presented in a fixed order; that is, participants would anticipate the identity of the next vowel. We found a consistent serial effect caused by C_{t-1} (N = 12; p = 0.0001, two-sided bootstrap, FDR corrected; **Fig. 3B**). Importantly, the amplitude of the DoG curve during fixed-order vowel production was similar to that during random-order (p = 0.307, two-sided bootstrap, FDR corrected; **Fig. 3C**), suggesting that serial learning of motor compensation was reduced due to different vowels, instead of uncertainty of next production.

<Figure 3 is about here>

Experiment 3: inter-syllabic learning in sentence production within and across word boundary

Experiments 1 and 2 revealed that preceding motor compensation drove a trial-to-trial update in the current trial. To further validate this finding, we modified and extended the serial-dependence paradigm into a controlled naturalistic-like sentential context, where we selectively applied auditory perturbation to a syllable and remained its subsequent syllable unperturbed. This extension allows us to test how compensation-driven serial learning interacts with sentential structures and influences normal production without auditory perturbation. We manipulated the auditory feedback during the continuous production of sentences at different positions of syntactic structures in two separate experiments (N = 15 and 13 in Experiment 3a and 3b, respectively). During sentence production, auditory pitch perturbation was randomly applied to either the second, third, or fourth syllable in each trial (**Fig. 4AC**). We quantified compensation in the perturbed syllable and its serial changes in the next syllable, which could be within or across the word boundaries (shown as grey rectangles in insets; **Fig. 4BD**). We showed that upward pitch perturbation drove a significant compensatory response around 200 ms after perturbation onset, compared with the pitch trace without perturbation (t(14) < -40.30, p < 0.009, cluster-based permutation test; **Fig. S2**). We extracted 200-300 ms mean pitch as compensation amount in the perturbed syllable and as serial-learning amount in its next syllable. Only when the two syllables formed a word, the

compensation in the preceding syllable influenced the production of the current syllable (right plot in **Fig. 4B:** $r = 0.09$, $p = 0.003$; left plot in **Fig. 4D**: $r = 0.22$, $p = 1 \times 10^{-11}$; Spearman correlation). The correlation coefficients were larger in the within-word-boundary condition than in the across-word-boundary condition (**Fig. 4B**: $p = 0.180$ and 0.026 for position 2 vs 4 and for position 3 vs 4, respectively; **Fig. 4D**: $p = 0.010$ and 0.0005 for position 2 vs 3 and for position 2 vs 4, respectively; one-sided bootstrap, FDR corrected). In contrast, auditory perturbation in the preceding syllable did not contribute to inter-syllabic learning (**Fig. S3**). These results strongly suggested that inter-syllabic learning persisted in the current vocal production even without any perturbation and compensation, and the serial learning was constrained by word boundary.

<Figure 4 is about here>

Discussion

We investigated how previous sensory error and motor compensation regulated trial-to-trial learning. We showed that motor compensation in the preceding trial influenced speech production in the next trial, whereas sensory errors did not have any effects. Crucially, this compensation-driven serial learning was modulated by linguistic contexts -- persisting across different vowels in a string of production and across vowels embedded within the word boundary. These results suggested that instantaneous sensorimotor learning was driven by motor compensation and was constrained by the linguistic system in continuous speech.

The study introduces a serial-dependence paradigm in the domain of speech production that dissociates the influence of previous sensory error and motor compensation on the current vocal production. Our results indicate that trial-to-trial learning predominantly arises from rapid updates in the feedforward control system, driven by prior motor compensation rather than sensory errors. The degree of motor compensation leads to immediate updates in the feedforward process that regulate changes in the compensatory responses to feedback perturbation (Experiments 1 and 2, **Fig. 2** and **3**). Moreover, the compensation-driven serial learning was also observed in the production of syllables without feedback perturbation (Experiment 3, **Fig. 4**). These evidence suggest that instantaneous plasticity in speech production emerges primarily through immediate updates in the feedforward control process, initiated by preceding compensatory responses. Using a novel serial dependency paradigm, this study extended sensorimotor learning in a relatively longer time scale[32,35,36] to trial-by-trial instantaneous plasticity.

The relationship between within-trial compensation and next-trial learning is not linear; it is more likely a DoG-shaped curve. Specifically, if the previous compensation is large, the serial changes in the current trial become smaller, suggesting a dynamic adjustment process during speech planning.[5,6,37] Furthermore, this process is also constrained by the similarity of production targets between trials. If the target in the next trial changes to a different vowel, the trial-to-trial learning becomes weaker (**Fig. 3C**). Rochet-Capellan et al. [38] tested whether sensorimotor learning during training transfers to new utterances with different sounds and demonstrated that it is highly dependent on the specific utterances. That is, the experience in the motor domain becomes less useful when the next utterance is different.

Lastly, the extension of compensation-driven serial learning into sentence production reveals that serial learning only exists within word boundary (**Fig. 4**). In Experiment 3, participants may simply read out aloud each visually presented word without syntactic processing. If this were true, inter-syllabic learning, as observed in Experiments 1 and 2, would occur between every two syllables. However, the observed significant effects of syntactic structures are inconsistent with the alternative 'reading-only' hypothesis. Moreover, the extensive training on the three 5-syllabic sentences ensured that participants were extremely familiar with the materials and retrieved the simple sentential structure at the visual presentation of the first word. All the experimental settings and positive results support the hypothesis regarding the role of syntactic constraints in compensation-driven serial learning.

The results of within-word-boundary inter-syllabic learning suggest that words serve as a central unit in speech planning. Motor compensation in the preceding syllable can be updated to the next within-word syllable. However, any compensation-based experience is not available for the syllables across words. Psycholinguistic studies, such as those by Levelt [39], propose that during speech production, words are retrieved from the mental lexicon as a whole unit, followed by phonological encoding into syllabic and phonetic representations. The entire course of speech production at the linguistic level would interact with sensorimotor processes via internal forward models and regulate feedforward processes at both featural and temporal domains,[11,13,40-44] to achieve efficient online control. The observed serial effects within word structure suggest that the programming and regulation of speech production are presumably in the size of the encoding unit within the word structure, which is consistent with the hypothesis of interaction between linguistic and sensorimotor systems in speech production control.[4]

The study showed that previous motor compensation can modulate the subsequent vocal pitch production; whether this compensation-driven serial learning can be generalized to other speech domains, such as formants[19] and intensity[11] that are controlled by distinct vocal apparatus and may be mediated by potentially different mechanisms, requires further investigation. Considering a broader linguistic context in speech motor control is challenging and relevant studies are rare.[45,46] In this study, we used pitch as a representative dimension and carefully selected different combinations of consonants and vowels in continuous speech to pioneer the investigation. Future studies should be carried out with caution and careful controls. For example, pitch contours in continuous speech can be potentially interfered with by consonantal perturbations[47] or sentence intonation[48]. During the production of non-sonorant consonants, the vibration of vocal folds will be affected, leading to a break and a carry-over rising or falling in the pitch contours.[49]

In summary, by implementing a novel serial-dependence paradigm in the domain of speaking, we found distinct functions of sensory error and motor compensation in sensorimotor learning in continuous speech. Motor compensation drives updates in the feedforward control process and regulates trial-to-trial learning that was constrained by linguistic contexts, while sensory error only contributed to the instant computation of compensation. These results suggest that the history of motor processes significantly contributes to sensorimotor learning and control. Systematic interplay among cognitive, motor, and sensory systems regulates behaviors.

Materials and methods

Participant

In five experiments, i.e., Experiments 1, 2a, 2b, 3a, and 3b, we recruited 14 participants (9 females, 18-27 years old, mean 21 years old), 17 participants (11 females, 19-32 years old, mean 23 years old), 12 participants (8 females, 20-27 years old, mean 22 years old), 15 participants (9 females, 19-25 years old, mean 22 years old), and 13 participants (7 females, 18-26 years old, mean 21 years old), respectively. They were college students with self-reported normal hearing and speech. All participants were right-handed and were native Chinese speakers. Each volunteer only participated in one experiment. Written informed consent was obtained from each participant before the experiment. The experimental procedures were approved by the Research Ethics Committee at the East China Normal University and New York University Shanghai.

Procedure in Experiment 1

Participants were asked to produce the vowel /a/ in a total of 960 trials over three days. Each trial began with a white fixation cross, displayed at the center of the screen for a randomized duration of 0.8–1.2 s (uniform distribution), signaling participants to be prepared for vocalization. When the cross turned green, participants were instructed to vocalize no more than 500 ms and sustain the /a/ sound with a steady and natural tone for 2.5 s until the cross disappeared. A pitch perturbation for a duration of 500 ms was randomly introduced 0.8–1.2 s after vocal onset. The direction (i.e., upward or downward) and magnitude (i.e., 60, 180, or 300 cents) of the pitch shift were randomly assigned across the 960 trials for each participant. At the end of each trial, participants received visual feedback: "Good," "Too slow," or "Too short." Participants completed 95.86% ± 0.02% (s.d.) of trials that met criteria. Each trial was followed by a silent interval with a random duration of 1–2 s (uniform distribution) before the next trial began.

Participants completed 320 trials in 10 blocks each day, with self-paced rest periods of at least 1 minute between blocks. At the beginning of each session, participants underwent a training block of 10 trials. The mean pitch from these trials was used to define the pitch range in the Audapter toolbox (see *Auditory feedback perturbation* below).

Procedure in Experiment 2a

The procedure for Experiment 2a was identical to Experiment 1, except participants were required to produce four different vowels without semantic meaning—啊 (/a1/), 咿 (/i1/), 呃 (/e1/), and 哦(/o1/)—randomly assigned in each trial. A white letter corresponding to the syllable was presented at the center of the screen, and upon turning green, participants were instructed to vocalize a vowel. Participants completed 94.82% ± 0.11% of trials that met the criteria.

Procedure in Experiment 2b

The procedure for Experiment 2b was identical to Experiment 2a, except the order of four vowels were presented in a fixed order, i.e., /i1/-/a1/-/e1/-/o1/ in every four trials so

that participants would not anticipate the next vowel production. Participants completed 97.55% ± 0.03 of trials that met the criteria.

Procedure in Experiment 3a

The procedure for Experiment 3a followed that of Experiment 1, with two differences. First, participants completed 300 trials in a single day. Second, they were required to produce sentences with the same syntactic structure rather than isolated vowels. In each trial, one of three five-syllable sentences (i.e., 妈妈撸猫咪 [ma1 ma1 lu1 mao1 mi1], 丫丫邀巫医 [ya1 ya1 yao1 wu1 yi1], 依依摸乌鸦 [yi1 yi1 mo1 wu1 ya1]) was randomly selected (uniform distribution), with each syllable appearing sequentially at the center of the screen for 500 ms, followed by a 500 ms blank screen. The whole duration for a sentence production was 5 s. During vocalization, a whole-utterance pitch perturbation (upward or downward perturbation at 60, 180, 300 cents, randomly assigned) was applied to the second, third, or fourth word during vocalization (uniformed distribution). In 20% of the sentences, no perturbation was applied. The speech materials and task were carefully chosen to avoid the effect of consonantal obstruction[50] and the effect of articulators' dynamics on pitch contour, respectively. Participants completed 94.44% ± 0.07% of trials that met the criteria.

Before the main experiment, participants completed a training session in which they produced each of the three sentences four times. They were explicitly instructed on the content and syntactic structure of the sentences. Additionally, they were instructed to vocalize each syllable within 500 ms of its appearance and to avoid coarticulation between words. Participants could repeat the training session upon request.

Procedure in Experiment 3b

The procedure for Experiment 3b was the same as Experiment 3a, except that the three sentences were replaced with another three with a different syntactic structure, i.e., 邀依琳摸鹰 [yao1 yi1 yi1 mo1 ying1], 拉妈妈腌鸭 [la1 ma1 ma1 yan1 ya1], and 约丫丫溜猫 [yue1 ya1 ya1 liu1 mao1], where the two-syllable noun was in the position of second and third character in the sentences. Participants completed 97.36% ± 0.03% of trials that met the criteria.

Auditory feedback perturbation

Participants' speech was recorded by a Shure beta 58A microphone that was placed 5 centimeters away from their mouths. The acoustic signals from the microphone were pre-amplified and then provided to an audio interface (MOTU Microbook IIc), collected at a 48k Hz sampling rate. The recorded speech was real-time manipulated (i.e., perturbation) in the audio interface, and the processed signal was split and sent both to the computer and to the headphones. The speech processing in the audio interface was performed using the Audapter toolbox.[51] The pitch of speech in each trial was shifted using the time-domain method. A 10-ms linear ramp was added at boundaries to attenuate possible abrupt jumps in the speech. A priori pitch range was set as +/-50 Hz relative to the mean pitch of a vocal /a/ for each participant. The rest of the parameters were set as default including algorithm, frame length, and cepstral liftering.

Acoustic analysis

In experiments 1 and 2, we first removed silence from the recordings of produced vowels in each trial using the *vad* function. Visual inspection was performed to make sure noise did not affect the voice activity detection. Pitch tracks (at a 100-Hz sampling rate) were measured using the Praat-based script ProsodyPro.[52] Default parameters were used. Perturbation onset in the pitch trace was obtained by a customized automatic algorithm combined with manual inspection. To convert hertz to cent scale, we defined baseline pitch by taking the mean of pitch trace in the -400 to -200 ms before perturbation onset. The pitch trace was transformed according to the formula: $cents = 1200 * \log_2(\frac{pitch}{pitch_{baseline}})$. A trial rejection was performed if the pitch trace was varied beyond its perturbation amount minus 100 cents (for downward-perturbation trials) or plus 100 cents (for upward-perturbation trials). To extract perturbation and compensation amounts for DoG-curve fitting and linear mixed effect model, the pitch trace in the 50-80 ms and the 150-250 ms windows were averaged, respectively, for each trial.[21] Compensation amounts for all trials and all participants were corrected by subtracting the grand average. This correction ensures that the DoG-curve fitting starts from zero, allowing its amplitude to be accurately measured and compared across experiments.

In Experiment 3, we applied the same pipeline except that only the perturbed syllable and its subsequent syllable were extracted to measure their pitch trace. Based on the visual inspection of pitch contours (**Fig. S2**), Mandarin tone 1 produced in isolation was characterized by a high-level tone and showed a pitch contour that typically involved a slight rising intonation at the onset, lasting around 0-150 ms, before stabilizing into a steady high pitch.[53] Therefore, we calculated the compensation amount in the perturbed syllable and serial-learning amount in the following syllable using mean pitch in the 200-250 ms time window, with baseline pitch in the 150-200 ms time window.

In all experiments, we analyzed signalOut, the speech output routed through the sound card to earphones, which incorporated the pitch perturbation. This approach is standard in perturbation studies e.g., [54]. A comparison of compensation amounts derived from signalIn and signalOut in Experiment 1 showed consistency between the two signals across participants (**Fig. S4**).

Statistical analysis of compensatory responses

For participant-level analysis, trials with the same perturbation condition were averaged, regardless of vowel categories in Experiments 2 and 3. To assess whether there was a compensatory response to perturbation, a two-sided cluster-based permutation test[55] was used to test averaged pitch traces against perturbation amount or between perturbed and normal conditions (**Fig. 1D** and **4B**). A one-sided one-sample t-test was used to test motor compensation extracted from time of interests against 0 (**Fig. 1E)**. Furthermore, to the main effect of P_{t-1} on C_t, a Friedman's chi-square test was used for multiple condition comparison (**Fig. 2D** and **3C**)**.** Spearman correlation was used to find the relation between C_{t-1}/P_{t-1} and inter-syllabic learning within and across syntactic boundaries (**Fig. 4C-E** and **S3**). To compare the correlation coefficient in within- and across-boundary conditions, a bias-corrected and accelerated bootstrap was used.[56]

Statistical analysis of trial-to-trial learning effect

The serial dependence paradigm quantified the similarity between auditory perturbations in two consecutive trials, i.e., P_{t-1}-P_t, and motor responses in the preceding one trial, i.e., C_{t-1}, to access their serial effect on the current motor compensation, i.e., C_t. To explore the serial effect of sensory error on C_t, C_t was expressed as a function of the difference between P_{t-1} and P_t. For positive values of this difference, the preceding perturbation P_{t-1} was higher than the perturbation of the current trial P_t; similarly, a positive compensation amount denoted a trial in which participants compensated upward in response to perturbation. Consequently, data points in the scatter plot (**Fig. 2BC**) that had x- and y-values in the same sign indicated that C_t was in the direction of P_{t-1}. In addition, to explore the serial effect of motor compensation on C_t, C_t was expressed as a function of the compensation amount in the preceding trial C_{t-1}. Following the same principle, the positive value of previous and current compensation, which had x- and y-values in the same sign, denoted a trial in which C_t was in the direction of C_{t-1}.

We assess the trial-to-trial learning effect by polling the compensation amount of all trials and fitting the DoG-shape curve, a well-found psychophysical function in previous serial-dependence studies.[32,36] The DoG was given as $y = x * a * w * c * e^{-(wx)^2}$, where *a* was the curve amplitude, *w* was the curve width, and *c* was the constant $\sqrt{2}/e^{-0.5}$. To assess serial effects of preceding sensory error, *x* was P_{t-1}-P_t, and width parameter *w* of the curve was constrained to a range of plausible values (*w* = 0.001-0.005, corresponding to the curve peak between 130 and 530 cents perturbation difference). To assess the serial effect of preceding motor compensation, *x* was C_{t-1}, and width parameter *w* of the curve was constrained to a range of plausible values (*w* = 0.019-0.071, corresponding to the curve peak between 10 and 36 perturbation pitch difference). The model fitting used the function *fmincon* with default parameters. We employed a bias-corrected and accelerated bootstrap to evaluate the significance of the DoG models and to compare DoG models between conditions. We applied a fitted DoG model to the pooled group data and generated a permutation distribution for the DoG-curve amplitude parameter.[57,58] This permutation process was repeated 2,000 times. For a one-sided comparison of curve amplitude, if the data population in one condition was greater than A% of the data population in the other condition, the significance level was (100A + 1)/2,001. For a two-sided comparison, the significance level was (200A + 1)/2,001. A false discovery rate (FDR) correction was applied.

Generalized linear mixed-effect model

To quantitatively test the relations of C_t to P_t, P_{t-1} and C_{t-1}, we built a generalized linear mixed-effect models (**Fig. 2E** and **3D**). The model assumed that P_t, P_{t-1} and C_{t-1} together influenced C_t. The function 'fitglme' was performed and accounted for group-level fixed effects and random effects because of vocal pitch variation across participants. The model was specified as $C_{t\ ij} = f(\beta_{0j} + \beta_{1j}P_{t\ ij} + \beta_{2j}P_{t-1\ ij} + \beta_{3j}C_{t-1\ ij} + 1|subjects)$, where *i* indicated participant index and *j* indicated j^{th} trial. The regression coefficient of each predictor, i.e., β, was accessed by the *t*-statistics against the null hypothesis test that the coefficient was equal to 0. FDR correction was applied.

Data and code availability: All data and code that support the findings of this study is available in Mendeley Data: 10.17632/pybdtwxvzz.2.

Acknowledgments: We thank Yuan Xie and Bao Hong for their thoughtful comments on previous versions of the manuscript, Tingting Wang and Huixuan Liu for assisting in collecting data. This study is 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 to X.T.

Author contributions: Y.L., X.T., and J.C. designed the experiment. Y.L., W.T., and Z.X. collected the data, Y.L., Z.X., A.X., and W.T., performed the analyses, Y.L., and X.T. drafted, reviewed, and corrected the manuscript. X.T. and J.C. supervised the project.

Declaration of interests: The authors declare no competing financial interests.

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the authors used ChatGPT to proofread texts. After using this tool or service, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Supplemental information: Document S1. Figures S1-S4.

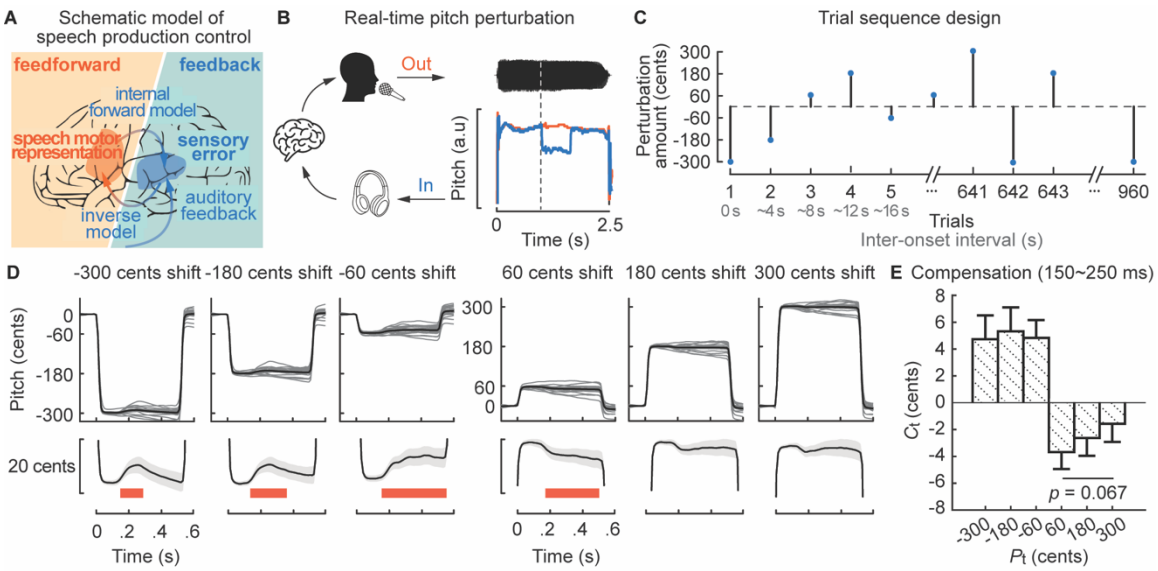

Figure 1. Motor compensation to pitch perturbation in the current trial. A, Schematics of a processing model of speech production control. Feedback and feedforward processes interactively mediate the control of speech production. When auditory feedback is perturbed and is inconsistent with the predicted consequence from an internal forward model, sensory error is generated and triggers the inverse model in the feedback control. The output of the feedback control further updates the programming of speech-motor action to compensate for the auditory perturbation. The auditory perturbation and resultant motor compensation could leave traces in either feedback and/or feedforward control processes that mediate sensorimotor learning, even on a trial-by-trial basis. **B**, The trial procedure of pitch perturbation design. Real-time pitch perturbation during production of vowel /a/. The onset of perturbation is denoted by a dashed line. The pitch of auditory feedback sent to earphones (blue) is shifted from the original utterance recorded in the microphone (orange). **C**, The trial sequence of the experiment. Participants undergo 320 trials each day for three consecutive days. The perturbation amounts and directions are randomly applied to each trial. **D**, Pitch traces with upward and downward perturbations and resultant compensation. The bold black lines denote the averaged pitch trace across trials for each participant (thin grey curves). Lower panels show magnified pitch traces from the above plots. The red horizontal bars denote the times of the significant effects of compensation ($p < 0.05$). **E**, Temporal averaged compensation in the periods of interest (150-250 ms after the perturbation onset). In **D** and **E**, shaded areas and error bars denote the standard error of mean (SEM) across participants. *$p < 0.05$, **$p < 0.005$.

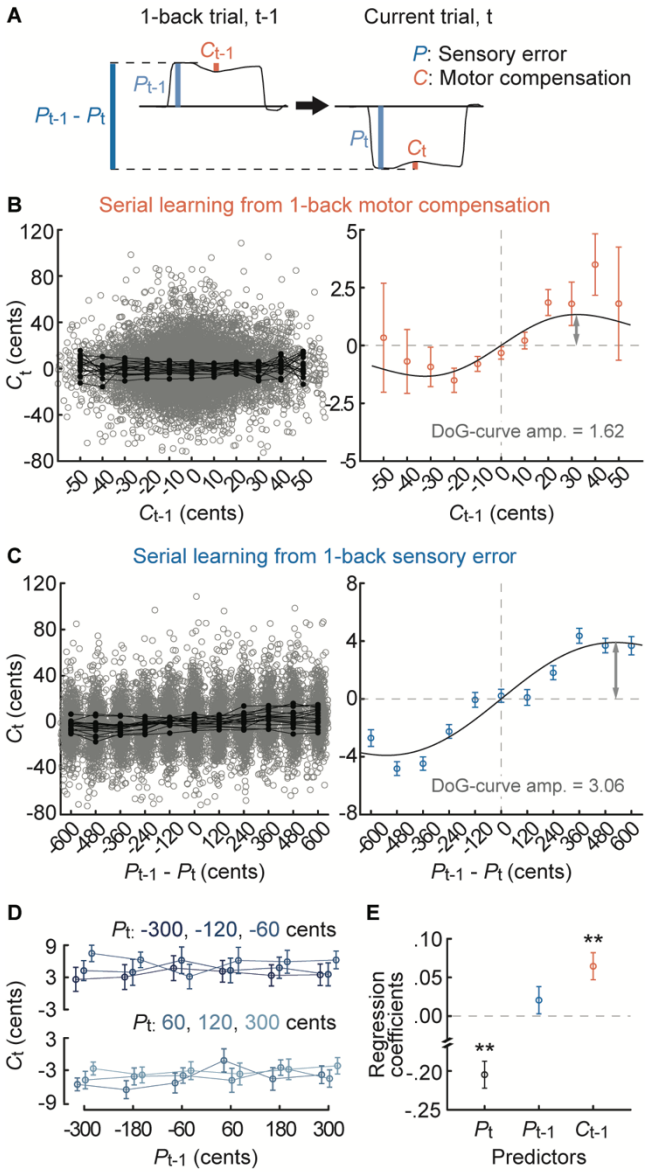

Figure 2. Serial changes of the immediately preceding trial on the compensatory responses of the current trial. A, Pitch trace of two adjacent trials. We measure the dependency of preceding sensory error and compensation on current speech production. In **B** and **C**, We plot compensation of the current trial, C_t, as a function of compensation in the immediately preceding trial, C_{t-1}, and as a function of the perturbation differences between the current and immediately preceding trial, P_{t-1} - P_t. In the left plots, each dot represents one trial and each black curve represents one participant. The right plots show the relationship between preceding and current trials after averaging across trials and participants. Color dots projecting to the same sign on the x and y axes indicate that C_t is in the same direction as C_{t-1} and P_{t-1} - P_t. **D**, C_t as a function of P_{t-1} does not show significant effects, suggesting the current compensation is not influenced by the perturbation amount of the immediately preceding trial. Error bars denote SEM across participants. **E**, The effects of current perturbation (P_t), preceding perturbation (P_{t-1}) and preceding compensation (C_{t-1}) on current compensation (C_t)

revealed by a generalized linear mixed-effect model. Regression coefficients for each predictor are extracted from the model. Current perturbation and preceding compensation exert opposite influences on the current compensation. Error bars denote SEM across trials. $^{**}p < 0.005$.

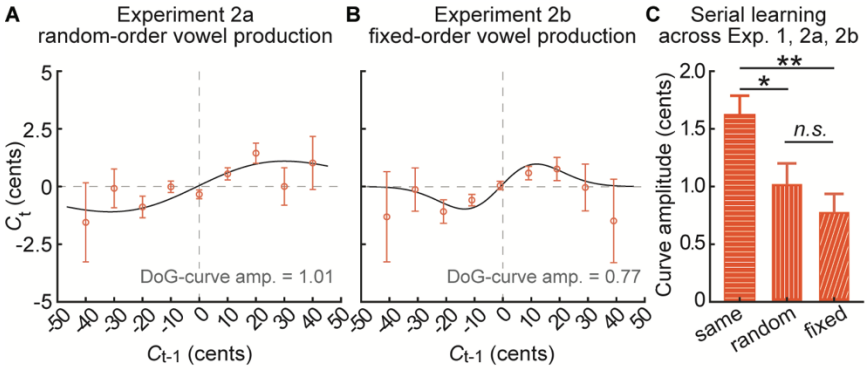

Figure 3. Serial effects across different vowel categories. A and **B**, C_t as a function of C_{t-1} when participants produce randomly-presented vowels (**A**) and when participants produce expectedly-presented vowels (**B**). In both experiments, significant serial changes to the current motor compensation are observed. **C**, The amount of compensation, characterized by the amplitude of the DoG curve shows differences between experiments 1, 2a, and 2b, where participants produce the same vowel, different vowels with uncertainty, and different vowels with certainty, respectively. The learning effects are weaker when producing different vowels, compared with that of the same vowels. Error bars denote one SD of the bootstrap distribution. *$p < 0.05$, **$p < 0.005$. See also Figure S1.

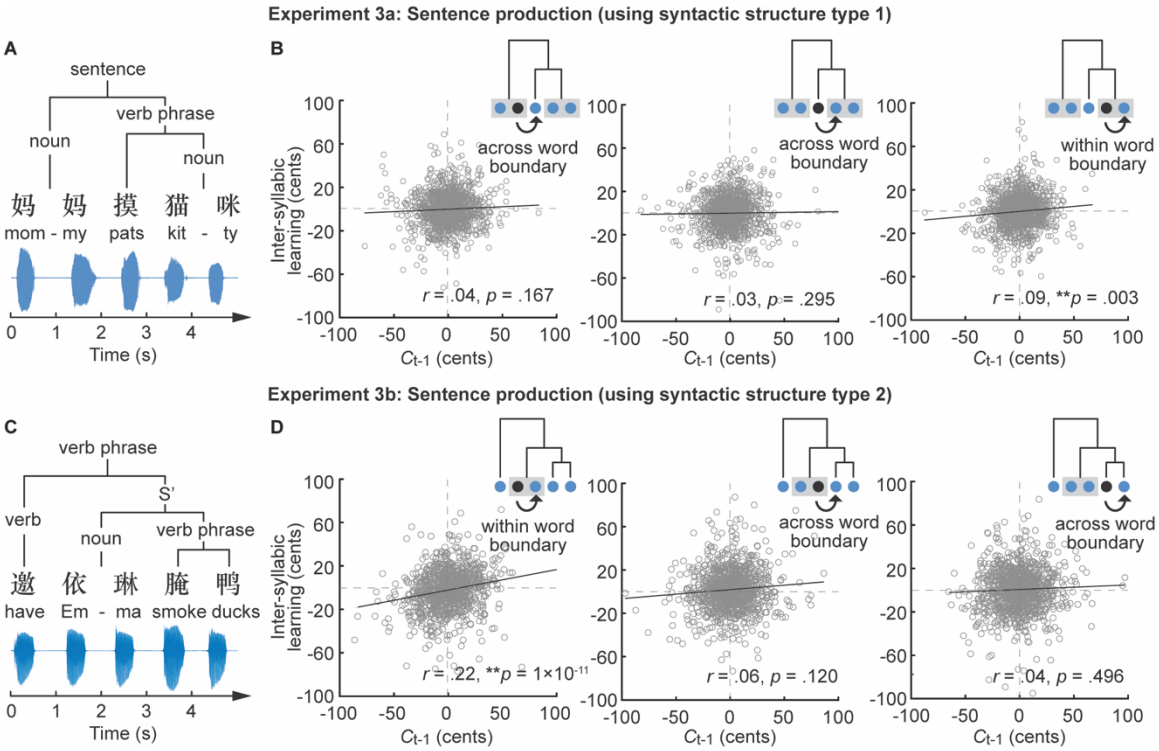

Figure 4. Inter-syllabic learning persists only within the word boundary. A and **C**, Experiment 3a and 3b. Participants are asked to produce three 5-syllabic sentences that have the same syntactic structure in each experiment. The position of the two-syllable noun is either at the end (Exp. 3a) or in the middle of the sentence (Exp. 3b). Each character is presented on the screen for 500 ms, followed by a blank for 500 ms. **B** and **D**, Relations between the compensation amount of the preceding syllable (x-axis) and serial-learning amount of the current syllable (y-axis) in Exp. 3a and 3b. Mean pitch averaged across 200-300 ms are separately extracted from the perturbed syllable (black dots in the insert plots) and the subsequent syllable (the dot pointed by a black arrow). The grey-shaped rectangle denotes the two-syllabic word. Only when the two syllables are within a word boundary, significant inter-syllabic learning is observed. **p < 0.005. See also Figure S2 and S3.

References

1. Rosenbaum, D.A. (2009). Human motor control (Academic press).
2. Ivanenko, Y.P., Poppele, R.E., and Lacquaniti, F. (2006). Motor control programs and walking. Neuroscientist *12*, 339-348. 10.1177/1073858406287987.
3. Olivier, E., Davare, M., Andres, M., and Fadiga, L. (2007). Precision grasping in humans: from motor control to cognition. Curr Opin Neurobiol *17*, 644-648. 10.1016/j.conb.2008.01.008.
4. Hickok, G. (2012). Computational neuroanatomy of speech production. Nat Rev Neurosci *13*, 135-145. 10.1038/nrn3158.
5. Houde, J.F., and Nagarajan, S.S. (2011). Speech production as state feedback control. Front Hum Neurosci *5*, 82. 10.3389/fnhum.2011.00082.
6. Guenther, F.H. (2016). Neural control of speech (Mit Press).
7. Tian, X., and Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. Front Psychol *1*, 166. 10.3389/fpsyg.2010.00166.
8. Kawato, M. (1999). Internal models for motor control and trajectory planning. Curr Opin Neurobiol *9*, 718-727. 10.1016/s0959-4388(99)00028-8.
9. Wolpert, D.M., and Ghahramani, Z. (2000). Computational principles of movement neuroscience. Nat Neurosci *3 Suppl*, 1212-1217. 10.1038/81497.
10. Schubotz, R.I. (2007). Prediction of external events with our motor system: towards a new framework. Trends in Cognitive Sciences *11*, 211-218. 10.1016/j.tics.2007.02.006.
11. Tian, X., Ding, N., Teng, X.B., Bai, F., and Poeppel, D. (2018). Imagined speech influences perceived loudness of sound. Nature Human Behaviour *2*, 225-234. 10.1038/s41562-018-0305-8.
12. Li, S., Zhu, H., and Tian, X. (2020). Corollary Discharge Versus Efference Copy: Distinct Neural Signals in Speech Preparation Differentially Modulate Auditory Responses. Cereb Cortex *30*, 5806-5820. 10.1093/cercor/bhaa154.
13. Li, Y., Luo, H., and Tian, X. (2020). Mental operations in rhythm: Motor-to-sensory transformation mediates imagined singing. PLoS Biol *18*, e3000504. 10.1371/journal.pbio.3000504.
14. Eliades, S.J., and Tsunada, J. (2018). Auditory cortical activity drives feedback-dependent vocal control in marmosets. Nat Commun *9*, 2540. 10.1038/s41467-018-04961-8.
15. Niziolek, C.A., Nagarajan, S.S., and Houde, J.F. (2013). What does motor efference copy represent? Evidence from speech production. J Neurosci *33*, 16110-16116. 10.1523/JNEUROSCI.2137-13.2013.
16. Chang, E.F., Niziolek, C.A., Knight, R.T., Nagarajan, S.S., and Houde, J.F. (2013). Human cortical sensorimotor network underlying feedback control of vocal pitch. Proc Natl Acad Sci U S A *110*, 2653-2658. 10.1073/pnas.1216827110.
17. Lametti, D.R., Nasir, S.M., and Ostry, D.J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. J Neurosci *32*, 9351-9358. 10.1523/JNEUROSCI.0404-12.2012.
18. Larson, C.R., Altman, K.W., Liu, H., and Hain, T.C. (2008). Interactions between auditory and somatosensory feedback for voice F0 control. Exp Brain Res *187*, 613-621. 10.1007/s00221-008-1330-z.
19. Tang, D.L., Parrell, B., and Niziolek, C.A. (2022). Movement variability can be modulated in speech production. J Neurophysiol *128*, 1469-1482. 10.1152/jn.00095.2022.
20. Houde, J.F., and Jordan, M.I. (1998). Sensorimotor adaptation in speech production. Science *279*, 1213-1216. 10.1126/science.279.5354.1213.
21. Hantzsch, L., Parrell, B., and Niziolek, C.A. (2022). A single exposure to altered auditory feedback causes observable sensorimotor adaptation in speech. Elife *11*. 10.7554/eLife.73694.
22. Tourville, J.A., and Guenther, F.H. (2011). The DIVA model: A neural theory of speech acquisition and production. Lang Cogn Process *26*, 952-981. 10.1080/01690960903498424.
23. Parrell, B., and Houde, J. (2019). Modeling the Role of Sensory Feedback in Speech Motor Control and Learning. J Speech Lang Hear Res *62*, 2963-2985. 10.1044/2019_JSLHR-S-CSMC7-18-0127.
24. Kim, K.S., Gaines, J.L., Parrell, B., Ramanarayanan, V., Nagarajan, S.S., and Houde, J.F. (2023). Mechanisms of sensorimotor adaptation in a hierarchical state feedback control model of speech.

PLoS Comput Biol *19*, e1011244. 10.1371/journal.pcbi.1011244.
25. Manes, J.L., Bullock, L., Meier, A.M., Turner, R.S., Richardson, R.M., and Guenther, F.H. (2024). A neurocomputational view of the effects of Parkinson's disease on speech production. Frontiers in Human Neuroscience *18*. 10.3389/fnhum.2024.1383714.
26. Tang, D.-l., Parrell, B., Beach, S.D., and Niziolek, C.A. (2025). The Brain's Sensitivity to Sensory Error Can Be Modulated by Altering Perceived Variability. The Journal of Neuroscience *45*, e0024242024. 10.1523/jneurosci.0024-24.2024.
27. Sitek, K.R., Mathalon, D.H., Roach, B.J., Houde, J.F., Niziolek, C.A., and Ford, J.M. (2013). Auditory cortex processes variation in our own speech. PLoS One *8*, e82925. 10.1371/journal.pone.0082925.
28. Niziolek, C.A., and Parrell, B. (2021). Responses to Auditory Feedback Manipulations in Speech May Be Affected by Previous Exposure to Auditory Errors. J Speech Lang Hear Res *64*, 2169-2181. 10.1044/2020_JSLHR-20-00263.
29. Kawato, M., Furukawa, K., and Suzuki, R. (1987). A hierarchical neural-network model for control and learning of voluntary movement. Biol Cybern *57*, 169-185. 10.1007/BF00364149.
30. Guenther, F.H. (2006). Cortical interactions underlying the production of speech sounds. J Commun Disord *39*, 350-365. 10.1016/j.jcomdis.2006.06.013.
31. Raharjo, I., Kothare, H., Nagarajan, S.S., and Houde, J.F. (2021). Speech compensation responses and sensorimotor adaptation to formant feedback perturbations. J Acoust Soc Am *149*, 1147. 10.1121/10.0003440.
32. Fischer, J., and Whitney, D. (2014). Serial dependence in visual perception. Nat Neurosci *17*, 738-743. 10.1038/nn.3689.
33. Cicchini, G.M., Mikellidou, K., and Burr, D.C. (2024). Serial Dependence in Perception. Annu Rev Psychol *75*, 129-154. 10.1146/annurev-psych-021523-104939.
34. Burnett, T.A., Freedland, M.B., Larson, C.R., and Hain, T.C. (1998). Voice F0 responses to manipulations in pitch feedback. J Acoust Soc Am *103*, 3153-3161. 10.1121/1.423073.
35. Fritsche, M., Spaak, E., and de Lange, F.P. (2020). A Bayesian and efficient observer model explains concurrent attractive and repulsive history biases in visual perception. Elife *9*. 10.7554/eLife.55389.
36. Liberman, A., Fischer, J., and Whitney, D. (2014). Serial dependence in the perception of faces. Curr Biol *24*, 2569-2574. 10.1016/j.cub.2014.09.025.
37. Houde, J.F., and Chang, E.F. (2015). The cortical computations underlying feedback control in vocal production. Curr Opin Neurobiol *33*, 174-181. 10.1016/j.conb.2015.04.006.
38. Rochet-Capellan, A., Richer, L., and Ostry, D.J. (2012). Nonhomogeneous transfer reveals specificity in speech motor learning. J Neurophysiol *107*, 1711-1717. 10.1152/jn.00773.2011.
39. Levelt, W.J.M. (1989). Speaking (MIT press). 10.7551/mitpress/6393.001.0001.
40. Tian, X., and Poeppel, D. (2015). Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and MEG. J Cogn Neurosci *27*, 352-364. 10.1162/jocn_a_00692.
41. Tian, X., and Poeppel, D. (2013). The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. J Cogn Neurosci *25*, 1020-1036. 10.1162/jocn_a_00381.
42. Indefrey, P., and Levelt, W.J. (2004). The spatial and temporal signatures of word production components. Cognition *92*, 101-144. 10.1016/j.cognition.2002.06.001.
43. Hartsuiker, R.J., and Kolk, H.H. (2001). Error monitoring in speech production: a computational test of the perceptual loop theory. Cogn Psychol *42*, 113-157. 10.1006/cogp.2000.0744.
44. Nozari, N., and Novick, J. (2017). Monitoring and Control in Language Production. Current Directions in Psychological Science *26*, 403-410. 10.1177/0963721417702419.
45. Patel, R., Niziolek, C., Reilly, K., and Guenther, F.H. (2011). Prosodic adaptations to pitch perturbation in running speech. J Speech Lang Hear Res *54*, 1051-1059. 10.1044/1092-4388(2010/10-0162).
46. Patel, R., Reilly, K.J., Archibald, E., Cai, S., and Guenther, F.H. (2015). Responses to Intensity-Shifted Auditory Feedback During Running Speech. J Speech Lang Hear Res *58*, 1687-1694.

10.1044/2015_JSLHR-S-15-0164.
47. Ohala, J.J. (1974). A mathematical model of speech aerodynamics. Annual Report of the Institute of Phonetics University of Copenhagen *8*, 11-22.
48. Daneš, F. (1960). Sentence intonation from a functional point of view. Word *16*, 34-54.
49. Xu, Y. (1997). Contextual tonal variations in Mandarin. Journal of Phonetics *25*, 61-83. DOI 10.1006/jpho.1996.0034.
50. Hanson, H.M. (2009). Effects of obstruent consonants on fundamental frequency at vowel onset in English. J Acoust Soc Am *125*, 425-441. 10.1121/1.3021306.
51. Cai, S., Boucek, M., Ghosh, S.S., Guenther, F.H., and Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong/iau. Proceedings of the 8th ISSP, 65-68.
52. Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. (Laboratoire Parole et Langage, France).
53. Tseng, C.-y. (1981). An acoustic phonetic study on tones in Mandarin Chinese (Brown University).
54. Smith, D.J., Stepp, C., Guenther, F.H., and Kearney, E. (2020). Contributions of Auditory and Somatosensory Feedback to Vocal Motor Control. J Speech Lang Hear Res *63*, 2039-2053. 10.1044/2020_JSLHR-19-00296.
55. Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods *164*, 177-190. 10.1016/j.jneumeth.2007.03.024.
56. Efron, B., and Tibshirani, R.J. (1994). An Introduction to the Bootstrap (CRC press). 10.1201/9780429246593.
57. Lu, Y.H., Jin, P.Q., Ding, N., and Tian, X. (2023). Delta-band neural tracking primarily reflects rule-based chunking instead of semantic relatedness between words. Cerebral Cortex *33*, 4448-4458. 10.1093/cercor/bhac354.
58. Lu, Y., Jin, P., Pan, X., and Ding, N. (2022). Delta-band neural activity primarily tracks sentences instead of semantic properties of words. Neuroimage *251*, 118979. 10.1016/j.neuroimage.2022.118979.