Distinct neural signals in speech preparation differentially modulate auditory responses

Siqi Li[1,2,*], Hao Zhu[2,3,*] & Xing Tian[2,3#]

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

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

[3] Division of Arts and Sciences, New York University Shanghai, China, 200122

* Both authors contributed equally

Correspondence to:

Xing Tian

Email: xing.tian@nyu.edu

Telephone: 86-21-20595201

Abstract

Actions influence sensory processing in a complex way to shape behavior. For example, it has been hypothesized that during actions, a copy of motor signals—termed *corollary discharge* or *efference copy*—can be transmitted to sensory regions and modulate perception. Such motor-to-sensory transformation has been evident among animal species and is extended to human speech production and control. The inhibitory function of the motor copies has been supported by the suppression of sensory responses during action execution. However, the sole inhibitory function is challenged by mixed empirical observations as well as multifaceted computational demands for behaviors. Theories have been proposed that *corollary discharge* and *efference copy* may be two separate functional forms that are generated at different stages of intention, preparation, and execution during actions. We tested these theories using speech in which we can precisely control and quantify the course of action. Specifically, we hypothesized that the content in the motor signals available at distinct stages of speech preparation determined the nature of signals (*corollary discharge* vs. *efference copy*) and constrained their modulatory functions on auditory processing. In three electroencephalography (EEG) experiments using a novel delayed articulation paradigm, we found that preparation without linguistic contents suppressed auditory responses to all speech sounds, whereas preparing to speak a syllable selectively enhanced the auditory responses to the prepared syllable. A computational model demonstrated that a bifurcation of motor signals could be a potential algorithm and neural implementation to achieve the distinct functions in the motor-to-sensory transformation. These consistent results suggest that distinct motor signals are generated in the motor-to-sensory transformation and integrated with sensory input to modulate perception.

Key words: internal forward model; efference copy; corollary discharge; prediction; motor control; speech production

Introduction

Actions influence sensory processing in a complex way to shape behavior. For example, the theory of internal forward models (Kawato, 1999; Schubotz, 2007; Wolpert & Ghahramani, 2000) proposes that during actions, a copy of motor signals, independently coined as *corollary discharge* (*CD*) by Sperry (1950) and *efference copy* (*EC*) by von Holst & Mittelstaedt (1950), can be transmitted to sensory regions and serves as a predictive signal to modulate sensory processing and perception. The common presumption regarding the functions of *CD* and *EC* is that these motor-to-sensory transformation signals suppress sensory processing in given modalities of perceptual consequences induced by motor actions. Based on the inhibitory functions, various cognitive abilities and behaviors can be achieved, such as efficient motor control (Kawato, 1999; Miall & Wolpert, 1996; Wolpert & Ghahramani, 2000), stable visual perception (Ross et al., 2001; Sommer & Wurtz, 2006), fluent vocal and speech production and control (Guenther, 1995; Hickok, 2012; John F Houde & Nagarajan, 2011; Tian, 2010), self-monitoring and agency (Sarah-Jayne Blakemore & Decety, 2001; Desmurget et al., 2009; Grush, 2004). Such motor-to-sensory transformation mechanisms have been evident among animal species (Crapse & Sommer, 2008) and their neural pathways have been increasingly mapped out (Poulet & Hedwig, 2006; Schneider et al., 2014, 2018).

However, the advances in anatomical and functional evidence for the motor-to-sensory transformation bring discrepancies. For example, computationally, the stability of perception during visual saccade requires motor signals to *suppress* the processing of sensory feedback (Ross et al., 2001). On the other hand, the predictive nature of motor signals mediates the receptive field remapping and *enhances* the sensory and perceptual sensitivity (Mohr et al., 2003; Neuweiler, 2003). Empirically, in addition to the commonly observed action-induced-suppression, action-induced-enhancement has also been found (Eliades & Wang, 2005, 2008; Enikolopov et al., 2018; Flinker et al., 2010; Singla et al., 2017). Cognitively, higher-level cognitive functions, such as self-monitoring, require motor signals to suppress the feedback to indicate the consequences

of self-induced actions (Sarah-Jayne Blakemore & Decety, 2001; Desmurget et al., 2009; Grush, 2004). Whereas for working memory and mental imagery, positive neural representations are needed to establish the mental images (Ma & Tian, 2019; Mary Zarate et al., 2015; Tian, 2010; Tian et al., 2018, 2016; Tian & Poeppel, 2012, 2013). Clinically, the loss of contact with the external world in psychosis has been assumed with the malfunction of agency via broken inhibitory functions from the motor system (Ford & Mathalon, 2004). However, the positive symptoms, such as auditory hallucinations, require actively induced the specific perceptual representations without corresponding external stimulations (Waters et al., 2012). The mixed observations and competing functions in motor-to-sensory transformation necessitate a reconsideration of the theoretical framework.

The observed action-induced sensory modulation occurs mostly during or after the execution of actions. Arguably, the execution phase is the last stage along with the entire action dynamics that include at least the intention and preparation stages (Fig. 1). These early stages before execution are mediated by the upper-stream motor circuitries that could potentially provide different motor signals (Crapse & Sommer, 2008; Straka et al., 2018). The last execution stage could bundle all the available motor signals and yield the observed mixed results and competing functions. We hypothesize, similar to previous theoretical proposals (Crapse & Sommer, 2008; Straka et al., 2018), that copies of distinct motor signals are available to transmit to sensory regions at distinct temporal stages. Specifically, the *CD* is available right after the initiation of action dynamics -- the intention of movement, whereas the *EC* is available only after the development of a concrete movement plan -- the encoding of movement (Fig. 1).

More importantly, we hypothesize that the functions of the distinct motor signals are determined by their contents. Similar to the arguments that a single type of corollary discharge would be too simplified to reflect the complexity of the motor signals regarding their sources, targets and functional utilities (Crapse & Sommer, 2008), we specify the putative functions by referring the literal meanings of the two historical terms. Specifically, the *corollary discharge* is a discharge signal within the established motor-

to-sensory transformation pathways. It does not necessarily include any content information. Its function could be, as manifested in saccadic suppression (Ross et al., 2001) and speech-induced suppression (Houde et al., 2002), inhibition of all processes in the connected sensory regions, indicating the impending motor actions (illustrated as a blue-shaded arrow in Fig. 1). Whereas, the *efference copy* is an identical copy of motor signals that include detailed codes about actions. It is generated in a manner of one-to-one mapping in the motor-to-sensory transformation pathway. Its function could be, as indicated in the amplification of the mormyromast electroreceptors for electrolocation in electric fishes (Mohr et al., 2003) and priming the echo-sensitive neurons for echolocation in bats (Neuweiler, 2003; Schuller, 1979), selectively enhancing the sensitivity to reafferent (sensory feedback) caused by actions (illustrated as a red-shaded arrow in Fig. 1). That is, we specify distinct functions in the otherwise interchangeably used historical terms to reflect our hypothesis that the functional specificity of motor signals is constrained by their contents. Together with the hypothesis about the distinct dynamics of these motor signals, the updated theoretical framework may account for the mixed neural modulations and competing functions of motor-to-sensory transformation.

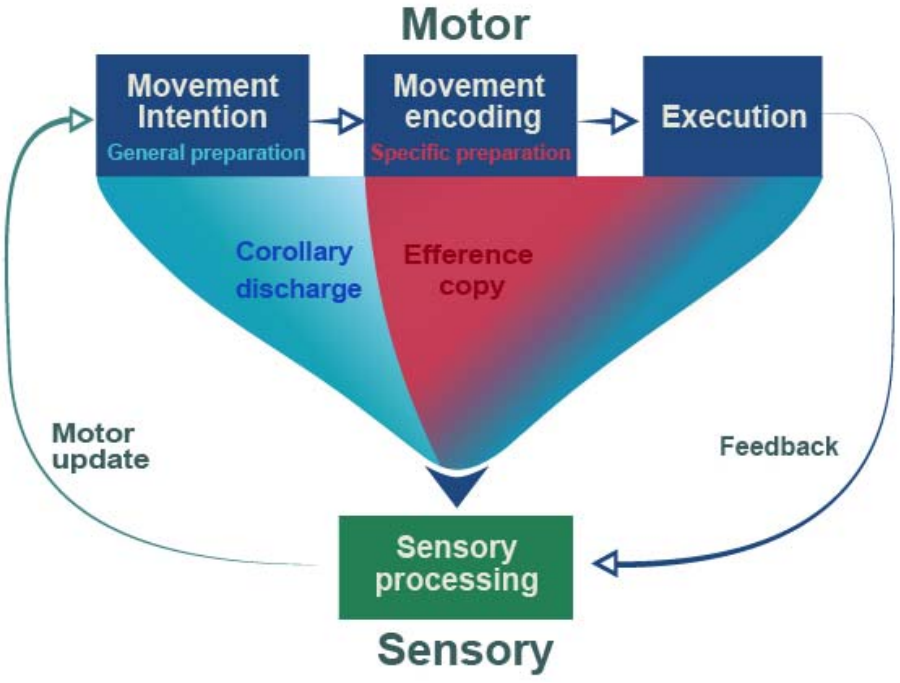

Figure 1. Schematics of proposed motor signals and their functions in the motor-to-sensory transformation. The intention and preparation stages before execution are mediated by the upper-stream motor circuitries that could potentially provide different motor signals. Copies of different motor signals are available to transmit to sensory regions at distinct temporal stages. The *corollary discharge* (*CD*) is a

discharge signal within the established motor-to-sensory transformation pathways and it would be available during the *general preparation* (*GP*) stage (in the movement intention phase). The *CD* does not necessarily include any content information. Its function could be inhibiting of all processes in the connected sensory regions, indicating the impending motor actions (blue-shaded arrow). Whereas the *efference copy* (*EC*) would be available during the *specific preparation* (*SP*) stage (in the movement encoding phase) — after the development of a concrete movement plan. Its function could be selectively modulating the neural responses to the prepared syllable sounds (red-shaded arrow). The last execution stage could bundle all available motor signals and yield the observed mixed results and competing functions.

In this study, we tested these hypotheses in the domain of human speech. The proposed functional specificity of the forward motor signals should be a canonical neural computation among animal species and across motor-related cognitive functions. However, because the nature of our hypotheses — the content information in different stages before action execution determines distinct functions of motor signals, the experimental manipulations on complex task requirements put a high demand on training animals. Therefore, we investigated these hypotheses with a novel delayed articulation paradigm using non-invasive human scalp electroencephalography (EEG) recordings. Participants prepared to speak according to different visual cues. When the cues were symbols, participants generally prepared the action of speaking without any linguistic information. According to our hypothesis, the *CD* would be available during this *general preparation* (*GP*) stage (as in the movement intention phase in Fig. 1) and would suppress neural responses to all sounds. When participants prepared to speak a syllable indicated by the written syllabic cues, the *EC* would be available during this *specific preparation* (*SP*) stage (as in the movement encoding phase in Fig. 1) and would selectively enhance the neural responses to the prepared syllable sounds.

Methods

Participants

A total of 16 volunteers (5 males; mean age = 23.13; age range, 19-31 years) participated in Experiment 1; 19 participants (5 males; mean age = 23.89; age range, 19-31 years) in Experiment 2, and 17 participants (4 males; mean age = 23.94; age range, 20-35 years) in Experiment 3. All participants were right-handed native Mandarin speakers from East China Normal University. All participants had normal hearing without neurological

deficits (self-reported). They received monetary incentives for their participation. Written informed consent was obtained from every participant. All protocols were approved by the institutional review board at New York University Shanghai.

Materials

Four audible syllables (/ba/, /ga/, /pa/, /ka/) and a 1k Hz pure tone with a duration of 400 ms were synthesized using the Neospeech web engine (www.neospeech.com) at a sampling rate of 44.1kHz in a male voice. All auditory stimuli were presented binaurally at 70 dB SPL, via plastic air tubes connected to foam earplugs (ER-3C Insert Earphones; Etymotic Research). A Shure beta 58A microphone was used to detect and record participants' vocalization. Materials were consistent throughout three experiments.

Procedure

Experiment 1: distinct preparation stages in a delayed-articulation task

Experiment 1 was an omnibus paradigm that included separate speech preparation stages before articulation. It aimed to prove the working principle of temporal segregating and inducing of *CD* and *EC* in different stages of action preparation. Auditory probes were introduced during each preparation stage to investigate the nature of the motor signals by testing how distinct preparation stages modulate the perceptual responses to the auditory probes.

We designed a delayed-articulation task. Participants were asked to produce a syllable according to visual cues after several possible stages of preparation. A trial started with a fixation cross of 500 ms, followed by a stage or a sequence of stages, each of which includes a visual cue that appeared in the center of the screen for a duration that jittered between 1100 ms to 1600 ms. Participants were instructed to make different preparations according to the cue. The visual cue was either meaningless symbols (#%) in yellow (blue in Fig.2*A* for better illustration) that did not contain any linguistic information (*general preparation, GP*) or a syllable in red that was identical to the one in the subsequent articulation task (*specific preparation, SP*). During the last 400 ms of each

preparation stage, either a 1k Hz pure tone or one of four auditory syllables (/ba/, /ga/, /pa/, /ka/) was presented to probe the modulatory function of motor preparatory signals. In *SP*, the auditory probe of syllables was either the same as or different from the visual cue, yielding two conditions -- auditory syllables were congruent with the visual information and hence the prepared syllable (*SPcon*) or incongruent (*SPinc*). After the offset of the sound, a blank with the duration jittered between 600 ms and 800 ms was presented and was followed by a syllable in green appeared in the center of the screen. Participants were asked to produce the syllable as fast and accurately as possible. The onset times of vocal responses were recorded to quantify the reaction time (Fig. 2*A*).

Visual cues were pseudorandomly paired and presented in a temporal order in a trial. For example, the green syllable articulation cue could immediately appear after the fixation (immediate vocalization without preparation, *NP*). The reaction time in *NP* trials served as a baseline behavioral responses of syllable production and compared with reaction times in other trials to quantify the effects of preparation behaviorally. The articulation cue could follow the general preparation cue (*GP*) or the specific preparation cue (*SP*). All cues could be presented in a sequence in a trial so that articulation task was performed after the general and specific preparation ($SP_{after\ GP}$), as illustrated in Fig. 2*A*. The time limits for articulation were set to 1500 ms, 1200 ms, and 1000 ms in *NP*, *GP*, and *SP* conditions, respectively. These manipulations were to eliminate any expectations and enforce preparation.

Moreover, in another type of trials where participants saw a white visual cue (**) without any linguistic information. No articulation green syllable cues were followed the white symbols in these trials. Participants only need to passively listen to the auditory probes without the requirement of action preparation or articulation (baseline listening without preparation, *B*). The *B* trials that had similar visual cues and auditory probes but without preparation yielded baseline auditory responses to quantify the neural modulation effects of preparation. Therefore, five types of trials (*NP*, *GP*, *SP*, $SP_{after\ GP}$, & *B*) were randomly presented in five blocks. In each block 64 trials were included, yielding a total of 320 trials in the experiment, with 60 trials for each type of the *NP*, *GP*, *SP*, and $SP_{after\ GP}$ trials

and 80 trials for *B*. The number of auditory probes in the *GP* and *SP* stages was 120 each, and each of the *SPcon* and *SPinc* conditions had 60 auditory probes.

Experiment 2: probabilistic auditory probes enforcing the general preparation

We varied the duration of visual cues to eliminate the temporal expectation of auditory cue onset time in Exp. 1. However, the auditory probes were always following the visual cue. This temporal association could grant participants a strategy that they could start to prepare after hearing the auditory probe. That is, the motor signals of interests were not induced throughout the preparation stages, which seriously dampened the modulation effects to the auditory probes, especially in the *GP* conditions as the null results in Exp.1. Experiment 2 aimed to control this confound by introducing trials that did not contain auditory probes. The mixed trials enforced participants to prepare to speak according to the visual cues even though they did not know what syllable to speak. This experimental manipulation increased the power to investigate the functions of *CD* during *GP*.

The experiment procedure was very similar to the one in Experiment 1, except that only the general preparation task (*GP*) was included. Half of the trials did not include the auditory probe (GP_{NS}), as illustrated in Fig. 3*A*. Such mixed trials enforced participants to prepare the final vocalization task based on the visual cues instead of auditory stimuli. The time limit for the articulation task was set to be 1500 ms for *NP*, 1200 ms for *GP* and GP_{NS} trials, respectively. Four types of trials (*NP*, *GP*, GP_{NS}, *B*) were randomly presented in four blocks. In each block, 64 trials were included, yielding a total of 384 trials in the experiment, with 96 trials in each of the *NP*, *GP*, GP_{NS} and *B*. Each of the *GP* and *B* conditions had 96 auditory probes.

Experiment 3: explicitly directing attention to auditory probes during preparation

Arguably, when preparation actions, attention is shifted to the perceptual consequences of actions. In Experiments 1 & 2, when preparing to speak, participants were likely to direct their attention to the sound that they were going to produce. Therefore, the observed

modulation effects on auditory probes could be induced by attention. However, because it is hard, if not impossible to completely wipe out attention, in this experiment we explicitly direct participants' attention to the auditory probes by a task related to the auditory probes. If the observations in Experiments 1 & 2 were caused by the attention, we should obtain similar results in this experiment. Otherwise, the results in previous experiments cannot be accounted for attention.

The experiment procedure is similar to Experiment 1, except that participants were required to identify the auditory probes. During the general preparation task, participants were asked to identify the upcoming auditory probe whether it was a syllable or a tone. During the specific preparation task, participants were asked to determine whether the visual cue and auditory probe were congruent or incongruent. Participants need to make the identification response within a time limit of 2000 ms (Fig. 4*A*).

Data analysis

Behavior data analysis

The reaction times (RTs) of the articulation task was calculated as the time lag between the onset of the green visual cue and the onset of vocalization. In Experiment 1, averaged RTs were obtained in each of the four trial types (*GP*, *SP*, $SP_{after\ GP}$, *NP*). The RT data was subject to a repeated-measures one-way ANOVA and a post-hoc Turkey Student *t-test* for pairwise comparisons. Behavioral data analysis in Experiment 2 was similar to Experiment 1 except that the reaction times from trials were averaged in each of the three trial types (*NP*, *GP*, GP_{NS}). The same statistical methods were applied. In Experiment 3, behavioral data analysis was identical to Experiment 1, averaged RTs were obtained in each of the four trial types (*GP*, *SP*, $SP_{after\ GP}$, *NP*).

EEG data acquisition and processing

Neural responses were recorded using a 32-channel active electrode system (Brain Vision actiCHamp; Brain Products) with a 1000 Hz sampling rate in an electromagnetic shield and sound-proof room. Electrodes were placed on an EasyCap, on which electrode

holders were arranged according to the 10-20 international electrode system. The impedance of each electrode was kept below 10 kΩ and the data were referenced online to the electrode of Cz and re-referenced offline to the average of all electrodes. Two additional EOG electrodes (HEOG and VEOG) were attached for monitoring ocular activity. The EEG data were acquired with Brain Vision PyCoder software (http://www.brainvision.com/pycorder.html) and filtered online between DC and 200 Hz with a notch filter at 50 Hz.

EEG data processing and analysis were conducted with customized Python codes, MNE-python (Gramfort et al., 2014), EasyEEG (Yang et al., 2018), and TTT toolboxes (Wang et al., 2019). For each participant's dataset, noisy channels were manually rejected during visual inspection. For each condition, epochs of responses to the auditory probe including a 200 ms pre-auditory probe period and an 800 ms post-auditory probe period were extracted. The epochs were band-pass filtered from 0.1 Hz to 30 Hz, and baseline corrected using the 200 ms pre-auditory probe period. Epochs with artifacts related to eye blinks and head movement were manually rejected. Epochs with peak-to-peak amplitude exceeded 100 μV were automatically excluded. To ensure data quality, epochs were excluded prior to analysis if they were contaminated by any residual noise. The remaining epochs were used to obtain the average event-related responses (ERP) in each condition. An average of 244 (*SD* = 42.3) epochs for each participant were included In Experiment 1, 208 (*SD*=22.0) epochs in Experiment 2, and 233 (*SD*=28.7) epochs in Experiment 3.

In Experiment 1, the global field power (GFP) -- the geometric mean across 32 electrodes -- was calculated separately for tones in three conditions (*GP*, *SP*, and *B*), and for the auditory probes of syllables in four conditions (*GP*, *SPcon*, *SPinc*, and *B*). Individual peak amplitudes and peak latencies for the N1 and P2 components in the GFP waveforms were automatically identified using the TTT toolbox in predetermined time windows of 90 to 110 ms and 190 to 210 ms, respectively (Wang et al., 2019). We visually verified that individually identified peaks by the toolbox were within the correct time windows in each participant. For the auditory probes of syllables, paired *t*-tests were carried out

between the auditory probe in *GP* and *B* conditions, as well as comparing *SPcon* and *SPinc* conditions to *B*, separately for the N1 and P2 components. For tones, repeated-measures one-way ANOVAs were conducted among the responses to the auditory probes in *GP*, *SP*, and *B*, separately for the N1 and P2 components.

In Experiment 2, EEG data analysis was similar to Experiment 1. For the auditory probes of syllables, paired *t*-tests were carried out between the auditory probe in *GP* and *B* conditions, separately for the N1 and P2 components. For the GP_{NS} condition, epochs were extracted around a similar post-stimulus time as *GP* condition. The same peak selection and ERP analysis methods were used to determine the exact peak latency to verify that no auditory responses were induced in GP_{NS}. Statistical methods were similar to Experiment 1. Both for the auditory probes of syllables and tones, paired *t*-tests were carried out between the auditory probe in *GP* and *B* conditions, separately for the N1 and P2 components.

In Experiment 3, EEG data processing was identical to Experiment 1. For the auditory probes of syllables, paired *t*-tests were carried out between the auditory probe in *GP* and *B* conditions, as well as comparing *SPcon* and *SPinc* conditions to *B*, separately for the N1 and P2 components. For tones, repeated-measures one-way ANOVAs were conducted among the responses to the auditory probes in *GP*, *SP*, and *B*, separately for the N1 and P2 components.

Modeling

To quantify the proposed distinctions between *CD* and *EC*, we built a two-layer neural network model to simulate the dynamics and modulation effects of motor signals on sensory processing (Fig. 5A). The upper layer represents the motor processing and the lower layer represents the auditory processing. Each layer includes multiple neurons that represent different syllables. (Only four nodes are drawn for illustration purposes.) Each neuron in the auditory layer is a rate-coded unit with synaptic depression. The updating of membrane potential is governed by Eq. 1.

$$\frac{dv_i(t)}{dt} = \tau\left\{g_i(1-v_i)\sum_j w_{ij}e_j - v_i[L + I(\sum_k o_k + n * m)]\right\} \qquad \text{(Eq. 1)}$$

The member potential of neuron *i* at the auditory layer, v_i, is updated according to the integration rate (time constant, τ) summing over three sources of input. The first input is an excitatory input from acoustic signals, e_j, via bottom-up connection strength w_{ij}. This bottom-up input drives the membrane potential to 1 (governed by the multiplier of 1-*v*). The second input is the leak with the fixed term *L*. The third input is the inhibition, which is the strength of *I* multiplied by the sum of two terms. One is the lateral inhibition that is the sum of output at time *t* from *k* units at the auditory layer. Another is the inhibition from the motor layer, *n*m*, which is specified next. The combination of the leak and inhibition drives the membrane potential towards 0 (as the term in the bracket is multiplied by -*v*). The fixed parameters are similar to those used in previous studies (Huber & O'Reilly, 2003; Ma & Tian, 2019).

The influences of motor signals were modeled as two sets of free parameters. The motor signals come from the same motor units but split into two sources. One source is that the activities of all motor units integrate into an interneuron that inhibits all neurons in the auditory layer. For simplification, the inhibition from each neuron in the motor layer is assigned as a unit value, *m*. The equivalent inhibition effects from the interneuron are the sum of *n* motor units, *n*m*. This motor source simulates the hypothesized function of *CD*. Another source is the direct modulation between the corresponding syllable in two layers. This motor signal is modeled as a gain control parameter, g_i, which increases the gain of the corresponding auditory unit to the excitatory input. This motor source simulates the hypothesized function of *EC*.

During the simulation of specific preparation (*SP*), only the prepared syllable in the motor layer is activated. This causes only one unit input to the interneuron and a specific gain modulation on the auditory neuron of the prepared syllable. During the simulation of general preparation (*GP*), because of the lack of linguistic information, the preparation induces weak activates in all motor neurons.

Results

Experiment 1: distinct preparation stages in a delayed-articulation task

In Experiment 1, participants were asked to speak a syllable after various stages of preparation. A repeated-measure one-way ANOVA on RTs found a significant main effect of preparation ($F(3,45) = 97.720$, $p < 0.0001$). Further paired *t*-tests revealed that the onset of articulation was consistently faster after preparation. Specifically, articulation after *GP* (mean RT of 609 ms) was faster than immediate vocalization without preparation (*NP*, mean RT of 633 ms), ($t(15) = 3.177$, $p < 0.01$). RTs were much shorter after *SP* (445 ms) than *NP* ($t(15) = 10.678$, $p < 0.0001$). RTs were shortest when they articulate after *GP* and *SP* in a row ($SP_{after\ GP}$: 422 ms) ($t(15) = 10.584$, $p < 0.0001$) (Fig. 2*B*). These behavioral results suggested that participants engaged in speech preparation.

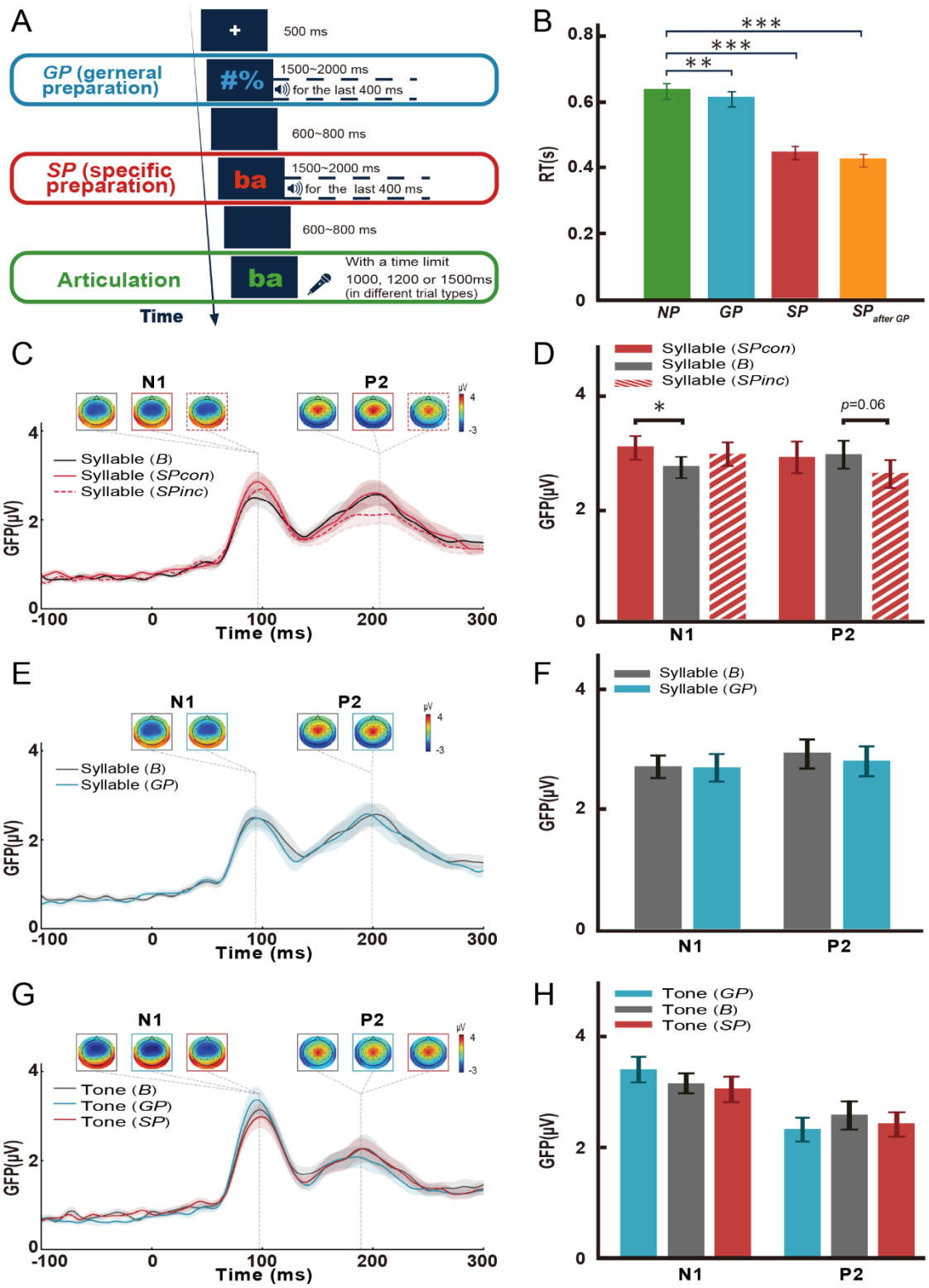

Figure 2. Experimental paradigm, behavioral, and ERP results of Experiment 1. A) Illustration of a sample trial that includes all preparation stages. Participants were asked to prepare to articulate a syllable according to visual cues that were either symbols (*general preparation*, *GP* – preparing to speak without knowing the content) or syllables in red (*specific preparation*, *SP* – preparing to speak the specific content). When a syllable in green appeared, participants were required to rapidly pronounce it. An auditory probe (a

1k Hz pure tone or a syllable sound) was presented during each preparation stage. Additional types of trials were included by randomly combined the preparation stages and the articulation tasks. For example, the articulation task can immediately follow the *GP* or can follow the *SP* without the preceding *GP*. Moreover, the articulation task can be presented without preparation (*no preparation, NP*). The articulation task was not required in the baseline passive listening trials (*B*). (Refer to Methods for all types of trials and conditions.) **B)** The speed of pronunciation measured as reaction time (RTs). Error bars indicate +- SEM. $^{**}p < 0.01$, $^{***}p < 0.001$. Faster articulation speed on *GP* and *SP* conditions than the *NP* condition. **C)** ERP time course and topographic responses for *SP* and *B* conditions. Individual peak amplitudes and peak latencies for the N1 and P2 GFP waveform were observed in each condition. The response topographies at each peak time are shown in colored boxes near each peak, using the same color-coding to represent each condition. The *SP* enhanced the N1 responses to the prepared syllables (*SPcon*). **D)** Mean GFP amplitudes across participants at N1 and P2 latencies for *SP* (red bars) and *B* (grey bars) conditions, respectively. *SP* enhanced the N1 responses to the prepared syllables (*SPcon*). Error bars indicate ± SEMs. Asterisks show the significance of post hoc *t-tests*, FDR-corrected for multiple comparisons ($^{*}p < 0.05$). **E)** ERP time course and topography responses for *GP* and *B* conditions show that no modulation effects of the *GP* on N1 and P2 responses. **F**) Mean GFP amplitudes across participants at the N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **E**. **G**) No modulation effects of *GP* and *SP* on the N1 and P2 responses to tones. **H**) Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones. The red bars depict the *SP* condition, the blue bars the *GP* condition and grey bar the *B* condition.

We further scrutinized the EEG neural responses to investigate the functions of motor signals during preparation. Paired *t*-tests were carried out between the auditory responses to the probes in the general preparation (*GP*) and specific preparation (*SP*) conditions, separately for the N1 and P2 components. In the *SP*, early neural responses of N1 were larger than that in *NP* when the auditory syllables were congruent with the specific preparation visual cues (*SPcon*) ($t(15) = -2.49$, $p = 0.025$). However, the effect was not significant in the later auditory responses of P2 ($t(15) = 0.248$, $p = 0.808$). The effects for the auditory syllables that were incongruent with the visual cue (*SPinc*) showed an opposite pattern. The effect in N1 was not significant ($t(15) = -1.48$, $p = 0.160$), whereas in P2 it is marginally significant ($t(15) = 2.024$, $p = 0.061$) (Fig. 2*D*). The neural response topographies were similar among *GP*, *SP*, and *B*, indicating that the observed effects were modulation solely on response magnitude but not the configuration of underlying neural sources (Fig. 2*C*). These results suggested that motor signals during specific preparation modulated the perceptual responses based on the content congruency.

In the *GP*, responses to auditory syllables were not different from the ones without preparation (*B*) in N1 ($t(15) = 0.07$, $p = 0.939$), nor in P2 ($t(15) = 1.070$, $p = 0.301$) (Fig. 2*F*). These results contrast with the ones obtained in the *SP*, presumably because motor signals with different natures were induced during distinct preparation stages. However,

the null results in *GP* were different from what we predicted -- the *corollary discharge* that was induced during *GP* would suppress auditory responses. That auditory probes were always at the last period of preparation stage would be a potential problem in this omnibus paradigm. That is, the general preparation could start toward the end of the stage so that the modulatory power of *corollary discharge* was significantly dampened. We addressed this potential problem in Experiment 2.

For the auditory probes of tones, repeated-measures one-way ANOVAs were conducted among three conditions (*GP*, *SP*, and *B*) for N1 and P2 separately. The effect was not significant both in the early auditory responses of N1 ($F(2,30) = 2.894$, $p = 0.07$) nor in the later auditory response of P2 ($F(2,30) = 2.111$, $p = 0.138$). These results showed that no modulation effects of motor signals on tones during either preparation stages (Fig. 2*H*). These results of tones contrasted with the results of auditory syllables, indicating the motor signals during preparation contained the task-related information. In summary, the results of Experiment 1 suggested that different motor signals were generated during distinct preparation stages and modulated perceptual neural responses based on the contents of signals.

Experiment 2: probabilistic auditory probes enforcing general preparation

The temporal association between the visual cues and auditory probes in *GP* would dampen the effects of *corollary discharge* and cause the null results in Experiment 1. In this experiment, we added trials without auditory probes during *GP* so that participants must prepare to speak according to the visual cues without linguistic information. The behavioral data showed a significant main effect of preparation ($F(2,36) = 105.101$, $p < 0.0001$). Further paired *t*-tests revealed that RTs were facilitated when participants performed *GP* with sound probe than immediate articulation *NP* (mean RT for *NP* 651 ms; *GP*, 580 ms; $t(18) = 10.534$, $p < 0.0001$). These results replicated the observations obtained in Experiment 1. More importantly, RTs were also significantly shorter when participants performed *GP* without sound probes than immediate articulation (mean RT for GP_{NS} 586 ms; $t(18) = 11.078$, $p < 0.0001$). These results suggested that participants performed the *GP* task according to the visual cues and ensured that *corollary discharge*

was available throughout the *general preparation* stage (Fig. 3*B*).

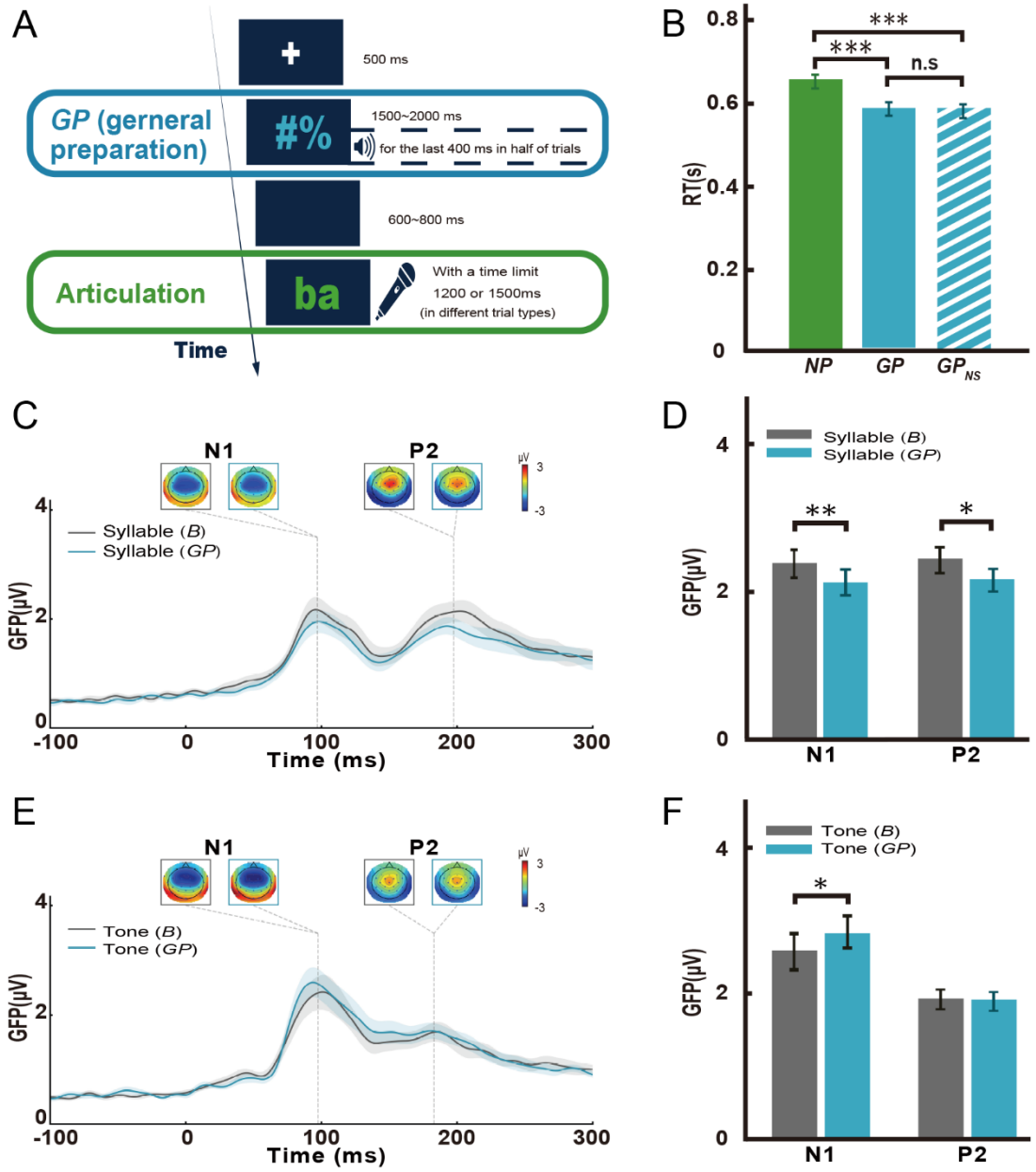

Figure 3. Experimental paradigm, behavioral, and ERP results for Experiment 2. A) Experiment 2 is similar to Experiment 1 except that participants performed the *GP* condition only. Half trials were without the auditory probes so that the mixed trials enforced participants to prepare to speak based on the visual cues without knowing the speech content. **B)** Facilitation in reaction time by the general preparation with (*GP*) or without sound probes (GP_{NS}). ***$p < 0.001$, *n.s*: not significant. **C)** ERP waveforms and topographic responses for *GP* and *B* conditions. The response topographies at each peak time are shown in colored boxes near each peak, using the same color-coding to represent each condition. Suppression in N1 and P2 responses to syllable sounds by the *GP* was observed. **D)** Mean GFP amplitude across participants at N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **C**. **E)** Enhancement in N1 responses to tones by the *GP*. **F)** Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones respectively. The blue bar depicts the *GP* condition and grey bar for the *B* condition.

For ERP responses to the auditory probes of syllables, paired *t*-tests revealed that the amplitude of early N1 response in *GP* was less than that in *B* ($t(18) = 3.406$, $p = 0.003$). The amplitude of later P2 response in *GP* was less than that in *B* ($t(18) = 2.240$, $p = 0.038$) (Fig. 3*D*). The topographies were consistent in both conditions (Fig. 3*C*). These results, obtained after presenting the auditory probes in a probabilistic manner and increasing the power of *corollary discharge*, were consistent with our hypothesis that *corollary discharge* that was induced during general preparation suppressed auditory responses.

Paired *t*-tests were also carried out on the auditory responses to the tones. Significantly larger N1 amplitude was revealed during *GP* compared with that in *B* ($t(18) = -2.397$, $p = 0.028$). The effect was not significant in the later auditory response of P2 ($t(18) = 0.193$, $p = 0.849$) (Fig. 3*F*). The enhancement in *GP* to the tones contrasted with the suppression results for the syllables, suggesting that the *corollary discharge* was motor-specific to actions -- speech production in this experiment. Violation of the goal of actions (e.g. pure tones that were not adapted to human vocal tracks and articulators), would create an error term and reverse the suppression effects. In summary, the results in Experiment 2 indicated the suppressive function of *corollary discharge*, which may be constrained by the task demand.

Experiment 3: explicitly directing attention to auditory probes during preparation

The modulation effects observed in Experiments 1 & 2 could be due to the shift of attention to the prepared speech sounds. However, it is hard, if not impossible, to disentangle motor preparation and attention. In this experiment, we explicitly instructed participants to identify the auditory probes during the *GP* and *SP* to examine whether the attentional effects differ from previous observations in Experiments 1 & 2. All participants successfully accomplished the identification task (accuracy of every participant was above 90%).

A repeated-measure one-way ANOVA on the articulation RTs revealed a significant main effect of preparation ($F(3,48) = 51.020$, $p < 0.0001$). A further paired *t*-test revealed

that RTs were shorter in *SP* than immediate articulation (*NP*: 621 ms, *SP*: 441 ms, $t(16) = 8.315$, $p < 0.0001$). RTs were shortest after having both *GP* and *SP* (437 ms) than immediate articulation ($t(16) = 7.234$, $p < 0.0001$). However, the RT difference between *GP* and *NP* was not significant (*GP*: 614 ms, $t(16) = 1.680$, $p = 0.135$) (Fig. 4*B*). Overall, the behavioral results were consistent with the findings in Experiment 1.

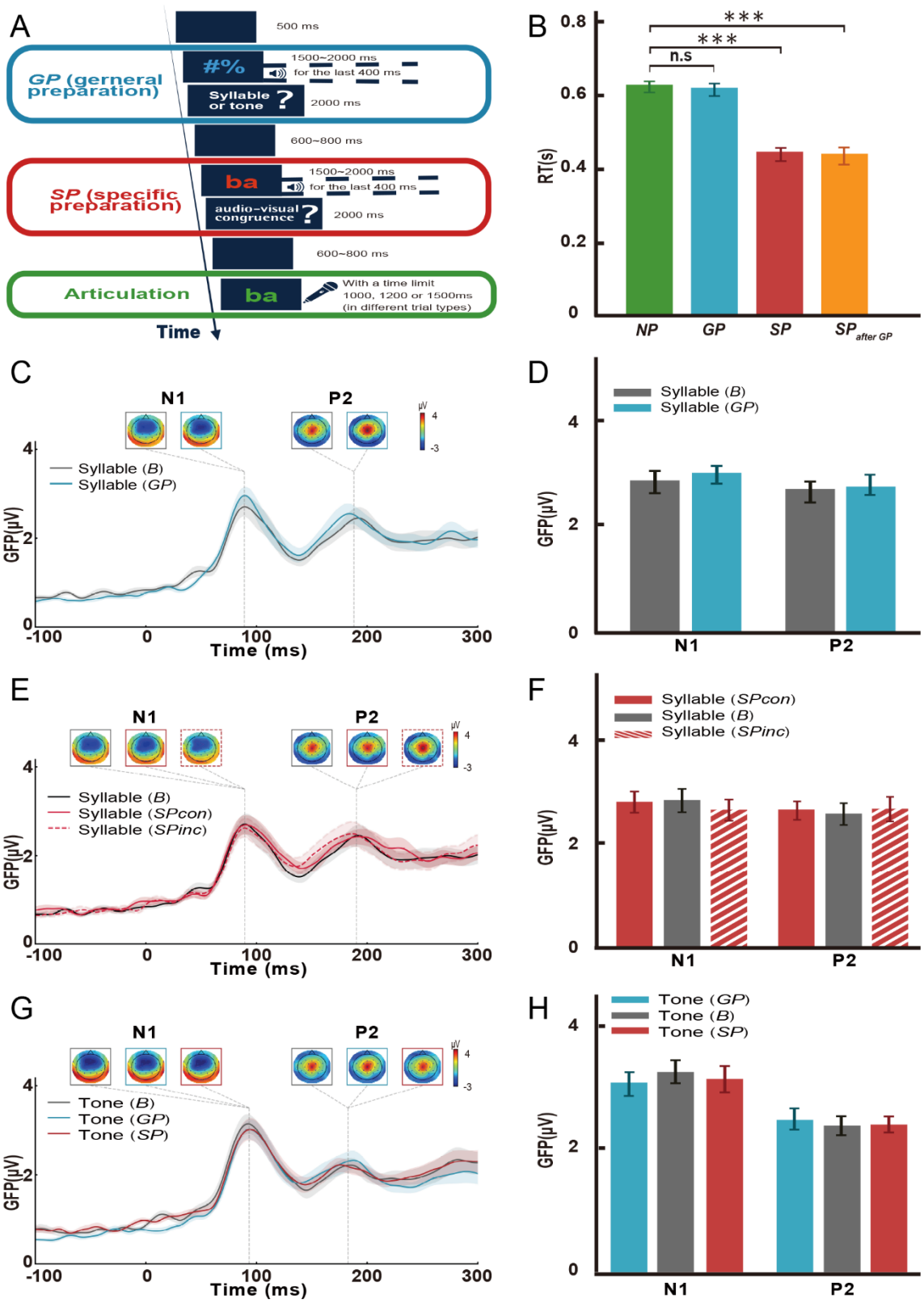

Figure 4. Experimental paradigm, behavioral, and ERP results for Experiment 3. A) Participants were explicitly instructed to identify the auditory syllables. During *GP*, participants were asked to identify the upcoming auditory probe whether it was a syllable or a tone. During *SP*, participants were asked to

determine whether the visual cue and auditory probe were congruent. **B)** Facilitation in reaction time by *SP*. ***$p < 0.001$. **C)** ERP time course and topographic responses for *GP* and *B* conditions. Individual peak amplitudes and peak latencies for the N1 and P2 GFP waveform were observed in each condition. The response topographies at each peak time are shown in colored boxes near each peak, using the same color-coding to represent each condition. The N1 and P2 responses for syllables in the *GP* condition are not significant. ***D***, Mean GFP amplitudes across participants at N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **C**. **E)** The effect was not significant in both the N1 and P2 response in the *SP* condition for syllables. **F**) Each bar represents the mean GFP amplitudes across participants at the N1 and P2 latencies for each condition to tones respectively. The red bars depict the *SP* condition and grey bar the *B* condition. **G**) No effects on the N1 and P2 responses to tones in both *GP* and *SP* conditions. **H**) Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones, the red bars depict the *SP* condition, the blue bars the *GP* condition and grey bar the *B* condition.

However, the effects in neural responses showed dramatic differences from the observations in previous experiments. Paired *t*-tests were conducted between conditions. For syllables, there was no significant difference between *GP* and *B* in either N1 or P2 response (N1: $t(16) = -1.052$, $p = 0.308$; P2: $t(16) = -1.068$, $p = 0.301$) (Fig. 4*D*). In the *SP*, the effect was not significant either in the N1 (*SPcon*: $t(16) = 0.136$, $p = 0.893$; *SPinc*: $t(16) = 1.162$, $p = 0.261$) or P2 (*SPcon*: $t(16) = -0.470$, $p = 0.645$; *SPinc*: $t(16) = -0.662$, $p = 0.517$) (Fig. 4*F*). A repeated-measures one-way ANOVA on responses to tones also did not reveal any significant results (N1: $F(2,32) = 0.535$, $p = 0.591$; P2: $F(2,32) = 0.154$, $p = 0.858$) (Fig. 4*H*). These null results after attentional manipulation clearly differed from the positive results obtained in motor preparation, suggesting that attention cannot account for the modulation effects observed in Experiments 1 & 2.

Discussion

We investigated the functions of motor signals along with the evolution of actions. With a novel delayed articulation paradigm in three electrophysiological experiments, we found that speech preparation at distinct stages differentially modulated auditory neural responses. When no linguistic information was available, the preparatory motor signals ubiquitously suppressed the early neural responses to all speech sounds. Whereas the preparatory motor signals generated based on a particular syllable enhanced the neural responses only to the prepared syllable. These modulatory functions in distinct directions along different stages of speech preparation suggest that granular motor signals with different natures were induced along the gradient of action dynamics.

Historically, *corollary discharge* and *efference copy* were proposed based on the observations of action execution. Arguably, execution is the ending output stage of an action, when presumably all possible motor signals are available. The lack of splitting the potentially complex motor signals may make the inhibitory functions of *CD* overwhelm other functions, yielding the well-established observation of action-induced sensory suppression. However, when considering the processing dynamics and signal contents in the hierarchy of the motor system, distinct motor signals are likely available at different stages (Crapse & Sommer, 2008; Straka et al., 2018) and exert distinct modulatory functions on the sensory systems. In this study, the dynamics and contents were experimentally isolated using a delayed articulation paradigm. This experimental manipulation revealed distinct modulatory functions of motor signals along with the evolution of actions, supporting the granular perspective of motor-to-sensory transformation.

Our behavioral and electrophysiological results cumulatively demonstrate that a type of motor signal can be generated during speech preparation even without any preparatory contents. Facilitation in articulation speed was observed after general preparation in both Exp. 1 and 2. Neural suppression in early auditory responses to syllable sounds was also observed in Exp. 2. These behavior and EEG results were consistent with immense literature about action-induced sensory suppression in both animal models (Crapse & Sommer, 2008; Eliades & Wang, 2008; Poulet & Hedwig, 2006; Schneider et al., 2018; Straka et al., 2018) and humans (Blakemore et al., 1998; Houde et al., 2002). Our results suggest that *CD* provides a uniform inhibitive function that suppresses sensory processing during the action. Moreover, our results reveal that *CD* is a generic form of motor signals that indicate the action, and can be available at the initial stage of action. This is consistent with the function of the *CD* on self-monitoring and agency (Desmurget et al., 2009; Kilteni et al., 2018; Tian et al., 2018).

The *CD* available during the general preparation enhanced the auditory responses to tones. These results suggest that *CD* is generated from and is constrained by the configuration of species' specific motor system. Although *CD* may not carry any specific content information, it is generated in the motor-to-sensory transformation pathways that adapt to

specific actions and reafferent sensory information. In our case, it is human speech — *CD* that is generated from the motor system controlling speech production is sent to and inhibits auditory cortices that represent the human speech sounds. The auditory system that represents pure tones may be relatively spare from inhibition, or its sensitivity maybe even relatively increase because tones are not adapted to human vocal tracks. Therefore, neural systems can separate the ex-afference sensory information (generated from external sources) from re-afference (feedback). These results are also consistent with the previous findings that showed relative increases in auditory responses when the speech feedback was substituted with non-speech sounds (Christoffels et al., 2011; Houde et al., 2002).

Comparing with the suppression of speech sounds during general preparation, the motor signals during the preparation of linguistic contents selectively modulated the auditory responses. That is, the motor signals during the specific preparation enhanced the auditory responses to the prepared syllables, whereas induced a mild suppression to unprepared syllables. These results suggest that *efference copy* carries specific content information, and selectively modulate the auditory system that represents the perceptual consequence of speaking. Our results are consistent with recent observations of action induced enhancement (Cao & Händel, 2019; Eliades & Wang, 2005, 2008; Enikolopov et al., 2018; Flinker et al., 2010; Ma & Tian, 2019; Singla et al., 2017; Tian & Poeppel, 2013; Tian et al., 2016).

The distinct directions of modulation effects on sensory processing at different preparation stages offer tantalizing hints suggesting that motor signals of distinct functions are available throughout the entire evolution of action. The *CD* can be available as soon as in the movement intention stage. It dissociates from specific actions that the system will engage, as we observed ubiquitous inhibition in auditory responses to all syllables. Moreover, the *CD* is probably independent of what motor effectors the actions would be executed by, as the auditory suppression was also observed by manual button-press (Bäß et al., 2008; Cao et al., 2017; Horváth et al., 2012). Furthermore, the *CD* is probably general inhibitory motor signals that are available across visual (Sommer & Wurtz, 2006), auditory (Poulet & Hedwig, 2006), and somatosensory (Sarah-J.

Blakemore et al., 1998) modalities.

On the contrary, the induction of *EC* requires a concrete action plan. The *EC* that contains specific action information would selectively enhance the perceptual responses to the same information that are contained in the motor signals, as we observed enhancement to the congruent syllable but not incongruent ones in the specific preparation stage. These results agree with the hypothesis of a one-to-one mapping between the motor and sensory systems (Ma & Tian, 2019; Tian & Poeppel, 2012, 2013; Tian et al., 2016). The enhancement effects of EC reflect the increased sensitivity to the congruent sensory representation, as compared to the incongruent stimuli (Eliades & Wang, 2008; Hickok et al., 2011; Ma & Tian, 2019). Because the concrete action information can be used to predict detailed perceptual consequences, the function of *EC* would be constrained by the established specific associations between motor and sensory systems, clearly contrasted with the ubiquitous inhibitory function of CD regardless of detailed sensorimotor correspondence.

The observations of dynamics and functional specificity of motor signals inspire upgrades of theories regarding sensorimotor integration and motor control. We put forward a tentative processing model in the framework of internal forward models with detailed temporal and functional features. *CD* is induced throughout the course of action. It exerts inhibitory functions on given sensory modalities via established motor-to-sensory transformation pathways. *EC* starts later when specific movement parameters are calculated. It increases the sensitivity to a given sensory token that relates to the results of the action via a detailed one-to-one mapping between the motor representation of the specific action and the linked sensory process of perceptual consequences. The complementary functions of *CD* and *EC* collaboratively enable self-monitoring and error detection/correction. The *CD* achieves the self-monitoring and agency by suppressing processes in a given sensory modality to indicate the non-specified perceptual consequence of one's actions. Whereas, the perceptual consequence of an action is sensitized by the *EC* so that the incrementally stronger *CD* when the concrete actions are carried out can precisely inhibit the sensory consequence and indicate possible errors of incorrect sensory feedback.

We quantify the proposed mechanism and potential neural implementation in a computational model (Fig. 5*A*). The copy of motor signals bifurcates. One branch has a direct one-to-one mapping and enhances the postsynaptic gain of the corresponding auditory unit (Ma & Tian, 2019). The increase of excitatory gain qualitatively equals to direct excitation but has sustaining effects as those during preparation. Another branch from all motor units activates an interneuron that inhibits all auditory units. During general preparation, activity from all motor units aggregately actives the interneuron that suppresses the neural responses to all syllables (Fig. 5*B*). When detailed information is available in the specific preparation stage, only one motor unit is activated. The excitatory effect from the motor unit outweighs its inhibitory effect and reverses the modulation into enhancement. During action execution, stronger signals from the local motor neurons inhibit the target sound, resulting in speech-induced suppression (e.g., Houde et al., 2002). That is, a parsimony model of motor signals bifurcation can account for the distinct functions observed in the action preparation and execution stages.

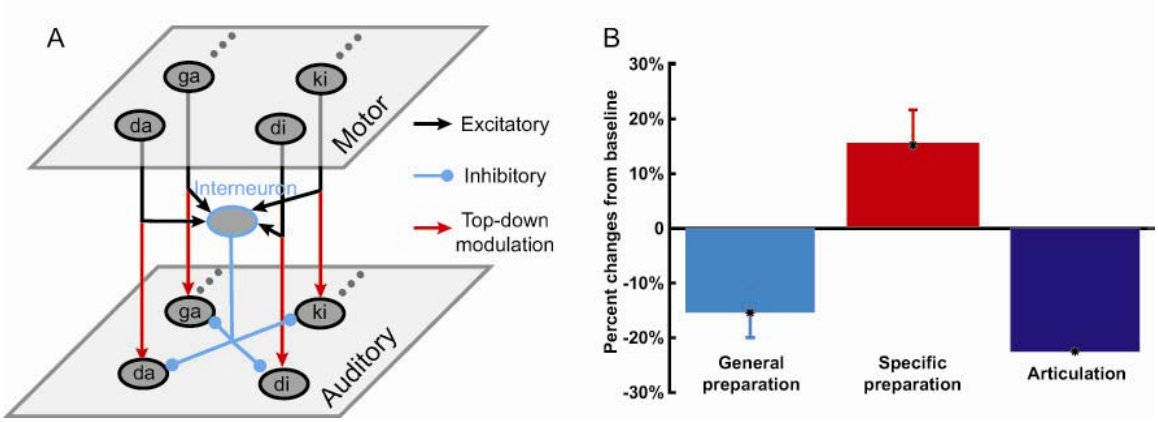

Figure 5. A neural network model of distinct motor signals and simulation results. A) Bifurcation of motor signals realizes distinct functions in a neural network model. A motor layer and an auditory layer include nodes that represent syllables. Each node is a rate-coded leaky-integrate-and-fire neuron. Signals from each motor unit split into two. One branch of the signals directly modulates the post-synaptic gain of the corresponding auditory unit, simulating the function of *EC* (the red line). The other branch of the signals accumulates and actives an interneuron that inhibits all auditory units, simulating the function of *CD*. **B)** Simulation results capture the modulation dynamics in speech preparation and execution. Bars are empirical data after converting into percent changes [(experimental condition - baseline)/baseline] and the stars are simulation results. The first two bars are modulation results of the general and specific preparations in Fig. 3D and Fig. 2D, respectively. The last bar is speech-induced suppression by averaging the effects from the left and right hemispheres in (Houde et al., 2002). Left, the lack of detailed information during the general preparation stage causes activation in all motor units and results in suppression of all auditory units. Middle, the detailed information available during the specific preparation stage activates a given motor unit and increases the sensitivity to the corresponding auditory unit, which yields an enhancement effect. Right, stronger inhibitory signals from a given motor unit during the execution of

speech strongly inhibit the corresponding auditory unit, resulting in the commonly observed speech-induced suppression.

The proposed mechanism can be tested in different sensory modalities in both humans and animal models. For example, the different timing and weighting of *CD* and *EC* could be realized by the onset of motor signals from different cortical areas in the motor hierarchy. In the visual domain, it could be the difference between the up-stream LIP for *CD* and downstream FEF for *EC* (Wang et al., 2016; Zirnsak et al., 2014). In the auditory domain, it could be the intention to speak in IPS (Tian, 2010) for the initialization of *CD* and frontal motor regions (including pre-motor, SMA, IFG) for *EC* (Tian et al., 2016). Moreover, motor signals are theorized to convey predictive signals to facilitate auditory perception and auditory-guided behaviors (Schneider et al., 2014). The different functions of motor signals could be manifested in the plasticity and modulation. In vision, the stabilization (temporal inhibition of visual processing) during saccade, remapping of the receptive field before saccade, and partial active receptive field during saccade (Sommer & Wurtz, 2006; Wang et al., 2016) could be caused by the interplay of distinct motor functions that modulate the visual processing. In audition, learning, self-monitoring of own articulation, differential manipulation of sensitive to auditory target, and speech error detection and correction (Hickok, 2012; Hickok et al., 2011; Houde, 1998; Liu & Tian, 2018; Tian & Poeppel, 2014) could be mediated by the interaction of distinct motor functions that modulate the auditory processing. The detailed functions and neural pathways about the proposed distinct motor signals could be further investigated and mapped out by electrophysiological, neuroimaging, and optic-genetic approaches (Poulet & Hedwig, 2006; Schneider et al., 2014, 2018; Sommer & Wurtz, 2006).

These results may offer insights about the cognitive neural mechanisms that mediated clinical and mental disorders. For example, our results may implicate a possible cause of auditory hallucinations from a perspective of internal monitoring and control. The normal population may use *EC* to internally induce auditory mental images and use the inhibitory function of *CD* to 'label' the source as internally self-generated. This interplay between *CD* and *EC* separates mental imagery from reality. However, patients suffering from auditory hallucinations may have intact *EC* to generate auditory mental images internally.

Whereas the inhibitory *CD* malfunctions (Tian & Poeppel, 2012; Yang et al., 2019). The intact enhancement function of *EC* generates auditory and speech representation based on internal stimulation of motor signals, but the lack of suppressive function of *CD* fails to label the internally generated sounds as self-generated. The internal prediction of a perceptual consequence, which has the same neural representation as an external perception, is erroneously interpreted as the result of external sources, which results in auditory hallucinations. Results in the current study support the hypothesis that two distinct motor signals are available to modulate perceptual responses, indicating their possible roles in speech monitoring and control, as well as the potential causes of auditory hallucinations.

Using a novel delayed articulation paradigm, we observed that distinct motor signals were generated in the motor-to-sensory transformation and integrated with sensory input to modulate perception during speech preparation. The content in the motor signals available at distinct stages of speech preparation determined the nature of signals—*corollary discharge* or *efference copy* and constrained their modulatory functions on auditory processing.

Acknowledgment

We thank Xingye Chen for her help in running experiments. This study was supported by the National Natural Science Foundation of China 31871131, the Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) 17JC1404104, and the Program of Introducing Talents of Discipline to Universities, Base B16018.

References

Bäß, P., Jacobsen, T., & Schröger, E. (2008). Suppression of the auditory N1 event-related potential component with unpredictable self-initiated tones: Evidence for internal forward models with dynamic stimulation. *International Journal of Psychophysiology*, *70*(2), 137–143. https://doi.org/10.1016/j.ijpsycho.2008.06.005

Blakemore, Sarah-J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640. https://doi.org/10.1038/2870

Blakemore, Sarah-Jayne, & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience*, *2*(8), 561–567. https://doi.org/10.1038/35086023

Cao, L., & Händel, B. (2019). Walking enhances peripheral visual processing in humans. *PLOS Biology*, *17*(10), e3000511. https://doi.org/10.1371/journal.pbio.3000511

Cao, L., Veniero, D., Thut, G., & Gross, J. (2017). Role of the Cerebellum in Adaptation to Delayed Action Effects. *Current Biology*, *27*(16), 2442-2451.e3. https://doi.org/10.1016/j.cub.2017.06.074

Christoffels, I. K., Ven, V. van de, Waldorp, L. J., Formisano, E., & Schiller, N. O. (2011). The Sensory Consequences of Speaking: Parametric Neural Cancellation during Speech in Auditory Cortex. *PLOS ONE*, *6*(5), e18307. https://doi.org/10.1371/journal.pone.0018307

Crapse, T. B., & Sommer, M. A. (2008). Corollary discharge across the animal kingdom. *Nature Reviews Neuroscience*, *9*(8), 587–600. https://doi.org/10.1038/nrn2457

Desmurget, M., Reilly, K. T., Richard, N., Szathmari, A., Mottolese, C., & Sirigu, A. (2009). Movement Intention After Parietal Cortex Stimulation in Humans.

Science, *324*(5928), 811–813. https://doi.org/10.1126/science.1169896

Eliades, S. J., & Wang, X. (2005). Dynamics of auditory–vocal interaction in monkey auditory cortex. *Cerebral Cortex*, *15*(10), 1510–1523.

Eliades, S. J., & Wang, X. (2008). Neural substrates of vocalization feedback monitoring in primate auditory cortex. *Nature*, *453*(7198), 1102–1106. https://doi.org/10.1038/nature06910

Enikolopov, A. G., Abbott, L. F., & Sawtell, N. B. (2018). Internally Generated Predictions Enhance Neural and Behavioral Detection of Sensory Stimuli in an Electric Fish. *Neuron*, *99*(1), 135-146.e3. https://doi.org/10.1016/j.neuron.2018.06.006

Flinker, A., Chang, E. F., Kirsch, H. E., Barbaro, N. M., Crone, N. E., & Knight, R. T. (2010). Single-Trial Speech Suppression of Auditory Cortex Activity in Humans. *Journal of Neuroscience*, *30*(49), 16643–16650. https://doi.org/10.1523/JNEUROSCI.1809-10.2010

Ford, J. M., & Mathalon, D. H. (2004). Electrophysiological evidence of corollary discharge dysfunction in schizophrenia during talking and thinking. *Journal of Psychiatric Research*, *38*(1), 37–46.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *Neuroimage*, *86*, 446–460.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*(3), 377–396. https://doi.org/10.1017/S0140525X04000093

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review, 102*(3), 594.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience, 13*(2), 135–145. https://doi.org/10.1038/nrn3158

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. *Journal of Cognitive Neuroscience, 24*(9), 1919–1931. https://doi.org/10.1162/jocn_a_00215

Houde, J. F. (1998). Sensorimotor Adaptation in Speech Production. *Science, 279*(5354), 1213–1216. https://doi.org/10.1126/science.279.5354.1213

Houde, John F, & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in Human Neuroscience, 5*, 82.

Houde, John F., Nagarajan, S. S., Sekihara, K., & Merzenich, M. M. (2002). Modulation of the Auditory Cortex during Speech: An MEG Study. *Journal of Cognitive Neuroscience, 14*(8), 1125–1138. https://doi.org/10.1162/089892902760807140

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: Temporal segregation through synaptic depression. *Cognitive Science, 27*(3), 403–430. https://doi.org/10.1207/s15516709cog2703_4

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current*

Opinion in Neurobiology, *9*(6), 718–727.

Kilteni, K., Andersson, B. J., Houborg, C., & Ehrsson, H. H. (2018). Motor imagery involves predicting the sensory consequences of the imagined movement. *Nature Communications*, *9*(1), 1617. https://doi.org/10.1038/s41467-018-03989-0

Liu, X., & Tian, X. (2018). The functional relations among motor-based prediction, sensory goals and feedback in learning non-native speech sounds: Evidence from adult Mandarin Chinese speakers with an auditory feedback masking paradigm. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-30399-5

Ma, O., & Tian, X. (2019). Distinct Mechanisms of Imagery Differentially Influence Speech Perception. *ENeuro*, *6*(5). https://doi.org/10.1523/ENEURO.0261-19.2019

Mary Zarate, J., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports*, *5*(1). https://doi.org/10.1038/srep11475

Miall, R. C., & Wolpert, D. M. (1996). Forward Models for Physiological Motor Control. *Neural Networks*, *9*(8), 1265–1279. https://doi.org/10.1016/S0893-6080(96)00035-4

Mohr, C., Roberts, P. D., & Bell, C. C. (2003). The Mormyromast Region of the Mormyrid Electrosensory Lobe. I. Responses to Corollary Discharge and Electrosensory Stimuli. *Journal of Neurophysiology*, *90*(2), 1193–1210. https://doi.org/10.1152/jn.00211.2003

Neuweiler, G. (2003). Evolutionary aspects of bat echolocation. *Journal of Comparative Physiology A*, *189*(4), 245–256. https://doi.org/10.1007/s00359-003-0406-2

Poulet, J. F. A., & Hedwig, B. (2006). The Cellular Basis of a Corollary Discharge. *Science*, *311*(5760), 518–522. https://doi.org/10.1126/science.1120847

Ross, J., Morrone, M. C., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends in Neurosciences*, *24*(2), 113–121. https://doi.org/10.1016/S0166-2236(00)01685-4

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

Schneider, D. M., Sundararajan, J., & Mooney, R. (2018). A cortical filter that learns to suppress the acoustic consequences of movement. *Nature*, *561*(7723), 391–395. https://doi.org/10.1038/s41586-018-0520-5

Schubotz, R. I. (2007). Prediction of external events with our motor system: Towards a new framework. *Trends in Cognitive Sciences*, *11*(5), 211–218.

Schuller, G. (1979). Vocalization influences auditory processing in collicular neurons of the CF-FM-bat,Rhinolophus ferrumequinum. *Journal of Comparative Physiology*, *132*(1), 39–46. https://doi.org/10.1007/BF00617730

Singla, S., Dempsey, C., Warren, R., Enikolopov, A. G., & Sawtell, N. B. (2017). A cerebellum-like circuit in the auditory system cancels responses to self-generated sounds. *Nature Neuroscience*, *20*(7), 943–950. https://doi.org/10.1038/nn.4567

Sommer, M. A., & Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature*, *444*(7117), 374–377. https://doi.org/10.1038/nature05279

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by

visual inversion. *Journal of Comparative and Physiological Psychology*, *43*(6), 482.

Straka, H., Simmers, J., & Chagnaud, B. P. (2018). A New Perspective on Predictive Motor Signaling. *Current Biology*, *28*(5), R232–R243. https://doi.org/10.1016/j.cub.2018.01.033

Tian, X. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00166

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

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: Linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*, 314. https://doi.org/10.3389/fnhum.2012.00314

Tian, X., & Poeppel, D. (2013). The Effect of Imagination on Stimulation: The Functional Specificity of Efference Copies in Speech Processing. *Journal of Cognitive Neuroscience*, *25*(7), 1020–1036. https://doi.org/10.1162/jocn_a_00381

Tian, X., & Poeppel, D. (2014). Dynamics of Self-monitoring and Error Detection in Speech Production: Evidence from Mental Imagery and MEG. *Journal of Cognitive Neuroscience*, *27*(2), 352–364. https://doi.org/10.1162/jocn_a_00692

Tian, X., Zarate, J. M., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, *77*, 1–12. https://doi.org/10.1016/j.cortex.2016.01.002

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Naturwissenschaften*, *37*(20), 464–476. https://doi.org/10.1007/BF00622503

Wang, Xiaolan, Fung, C. C. A., Guan, S., Wu, S., Goldberg, M. E., & Zhang, M. (2016). Perisaccadic Receptive Field Expansion in the Lateral Intraparietal Area. *Neuron*, *90*(2), 400–409. https://doi.org/10.1016/j.neuron.2016.02.035

Wang, Xuefei, Zhu, H., & Tian, X. (2019). Revealing the Temporal Dynamics in Non-invasive Electrophysiological recordings with Topography-based Analyses. *BioRxiv*, 779546.

Waters, F., Allen, P., Aleman, A., Fernyhough, C., Woodward, T. S., Badcock, J. C., Barkus, E., Johns, L., Varese, F., Menon, M., Vercammen, A., & Larøi, F. (2012). Auditory Hallucinations in Schizophrenia and Nonschizophrenia Populations: A Review and Integrated Model of Cognitive Mechanisms. *Schizophrenia Bulletin*, *38*(4), 683–693. https://doi.org/10.1093/schbul/sbs045

Wolpert, D. M., & Ghahramani, Z. (2000). Computational principles of movement neuroscience. *Nature Neuroscience*, *3*(11), 1212–1217. https://doi.org/10.1038/81497

Yang, F., Fang, X., Tang, W., Hui, L., Chen, Y., Zhang, C., & Tian, X. (2019). Effects and potential mechanisms of transcranial direct current stimulation (tDCS) on auditory hallucinations: A meta-analysis. *Psychiatry Research*, *273*, 343–349. https://doi.org/10.1016/j.psychres.2019.01.059

Yang, J., Zhu, H., & Tian, X. (2018). Group-level multivariate analysis in EasyEEG toolbox: Examining the temporal dynamics using topographic responses. *Frontiers in Neuroscience*, *12*, 468.

Zirnsak, M., Steinmetz, N. A., Noudoost, B., Xu, K. Z., & Moore, T. (2014). Visual space is compressed in prefrontal cortex before eye movements. *Nature*, *507*(7493), 504–507. https://doi.org/10.1038/nature13149

Figure captions

Figure 1. Schematics of proposed motor signals and their functions in the motor-to-sensory transformation. The intention and preparation stages before execution are mediated by the upper-stream motor circuitries that could potentially provide different motor signals. Copies of different motor signals are available to transmit to sensory regions at distinct temporal stages. The *corollary discharge* (*CD*) is a discharge signal within the established motor-to-sensory transformation pathways and it would be available during the *general preparation* (*GP*) stage (in the movement intention phase). The *CD* does not necessarily include any content information. Its function could be inhibiting of all processes in the connected sensory regions, indicating the impending motor actions (blue-shaded arrow). Whereas the *efference copy* (*EC*) would be available during the *specific preparation* (*SP*) stage (in the movement encoding phase) — after the development of a concrete movement plan. Its function could be selectively modulating the neural responses to the prepared syllable sounds (red-shaded arrow). The last execution stage could bundle all available motor signals and yield the observed mixed results and competing functions.

Figure 2. Experimental paradigm, behavioral, and ERP results of Experiment 1. A) Illustration of a sample trial that includes all preparation stages. Participants were asked to prepare to articulate a syllable according to visual cues that were either symbols (*general preparation, GP* – preparing to speak without knowing the content) or syllables in red (*specific preparation, SP* – preparing to speak the specific content). When a syllable in green appeared, participants were required to rapidly pronounce it. An auditory probe (a 1k Hz pure tone or a syllable sound) was presented during each

preparation stage. Additional types of trials were included by randomly combined the preparation stages and the articulation tasks. For example, the articulation task can immediately follow the *GP* or can follow the *SP* without the preceding *GP*. Moreover, the articulation task can be presented without preparation (*no preparation*, *NP*). The articulation task was not required in the baseline passive listening trials (*B*). (Refer to Methods for all types of trials and conditions.) **B)** The speed of pronunciation measured as reaction time (RTs). Error bars indicate +- SEM. $^{**}p < 0.01$, $^{***}p < 0.001$. Faster articulation speed on *GP* and *SP* conditions than the *NP* condition. **C)** ERP time course and topographic responses for *SP* and *B* conditions. Individual peak amplitudes and peak latencies for the N1 and P2 GFP waveform were observed in each condition. The response topographies at each peak time are shown in colored boxes near each peak, using the same color-coding to represent each condition. The *SP* enhanced the N1 responses to the prepared syllables (*SPcon*). **D)** Mean GFP amplitudes across participants at N1 and P2 latencies for *SP* (red bars) and *B* (grey bars) conditions, respectively. *SP* enhanced the N1 responses to the prepared syllables (*SPcon*). Error bars indicate ± SEMs. Asterisks show the significance of post hoc *t-tests*, FDR-corrected for multiple comparisons ($^{*}p < 0.05$). **E)** ERP time course and topography responses for *GP* and *B* conditions show that no modulation effects of the *GP* on N1 and P2 responses. **F**) Mean GFP amplitudes across participants at the N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **E**. **G**) No modulation effects of *GP* and *SP* on the N1 and P2 responses to tones. **H**) Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones. The red bars depict the *SP* condition, the blue bars the *GP* condition and grey bar the *B* condition.

Figure 3. Experimental paradigm, behavioral, and ERP results for Experiment 2. A) Experiment 2 is similar to Experiment 1 except that participants performed the *GP* condition only. Half trials were without the auditory probes so that the mixed trials enforced participants to prepare to speak based on the visual cues without knowing the speech content. **B)** Facilitation in reaction time by the general preparation with (*GP*) or without sound probes (GP_{NS}). $^{***}p < 0.001$, *n.s*: not significant. **C)** ERP waveforms and topographic responses for *GP* and *B* conditions. The response topographies at each peak

time are shown in colored boxes near each peak, using the same color-coding to represent each condition. Suppression in N1 and P2 responses to syllable sounds by the *GP* was observed. **D)** Mean GFP amplitude across participants at N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **C**. **E)** Enhancement in N1 responses to tones by the *GP*. **F)** Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones respectively. The blue bar depicts the *GP* condition and grey bar for the *B* condition.

Figure 4. Experimental paradigm, behavioral, and ERP results for Experiment 3. A) Participants were explicitly instructed to identify the auditory syllables. During *GP*, participants were asked to identify the upcoming auditory probe whether it was a syllable or a tone. During *SP*, participants were asked to determine whether the visual cue and auditory probe were congruent. **B)** Facilitation in reaction time by *SP*. *** $p < 0.001$. **C)** ERP time course and topographic responses for *GP* and *B* conditions. Individual peak amplitudes and peak latencies for the N1 and P2 GFP waveform were observed in each condition. The response topographies at each peak time are shown in colored boxes near each peak, using the same color-coding to represent each condition. The N1 and P2 responses for syllables in the *GP* condition are not significant. ***D***, Mean GFP amplitudes across participants at N1 and P2 latencies for *GP* (blue bars) and *B* (grey bars) conditions as observed in **C**. **E)** The effect was not significant in both the N1 and P2 response in the *SP* condition for syllables. **F**) Each bar represents the mean GFP amplitudes across participants at the N1 and P2 latencies for each condition to tones respectively. The red bars depict the *SP* condition and grey bar the *B* condition. **G**) No effects on the N1 and P2 responses to tones in both *GP* and *SP* conditions. **H**) Each bar represents the mean GFP amplitudes across participants at N1 and P2 latencies for each condition to tones, the red bars depict the *SP* condition, the blue bars the *GP* condition and grey bar the *B* condition.

Figure 5. A neural network model of distinct motor signals and simulation results. A) Bifurcation of motor signals realizes distinct functions in a neural network model. A motor layer and an auditory layer include nodes that represent syllables. Each node is a

rate-coded leaky-integrate-and-fire neuron. Signals from each motor unit split into two. One branch of the signals directly modulates the post-synaptic gain of the corresponding auditory unit, simulating the function of *EC* (the red line). The other branch of the signals accumulates and actives an interneuron that inhibits all auditory units, simulating the function of *CD*. **B)** Simulation results capture the modulation dynamics in speech preparation and execution. Bars are empirical data after converting into percent changes [(experimental condition - baseline)/baseline] and the stars are simulation results. The first two bars are modulation results of the general and specific preparations in Fig. 3D and Fig. 2D, respectively. The last bar is speech-induced suppression by averaging the effects from the left and right hemispheres in (Houde et al., 2002). Left, the lack of detailed information during the general preparation stage causes activation in all motor units and results in suppression of all auditory units. Middle, the detailed information available during the specific preparation stage activates a given motor unit and increases the sensitivity to the corresponding auditory unit, which yields an enhancement effect. Right, stronger inhibitory signals from a given motor unit during the execution of speech strongly inhibit the corresponding auditory unit, resulting in the commonly observed speech-induced suppression.

Figures

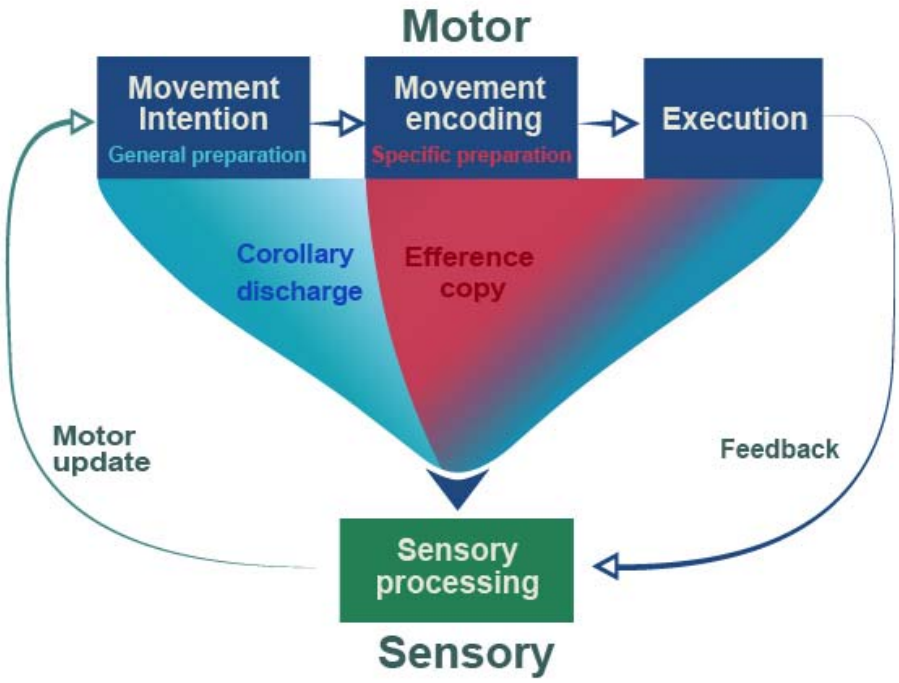

Figure 1. Schematics of proposed distinct motor signals and their functions in the motor-to-sensory transformation.

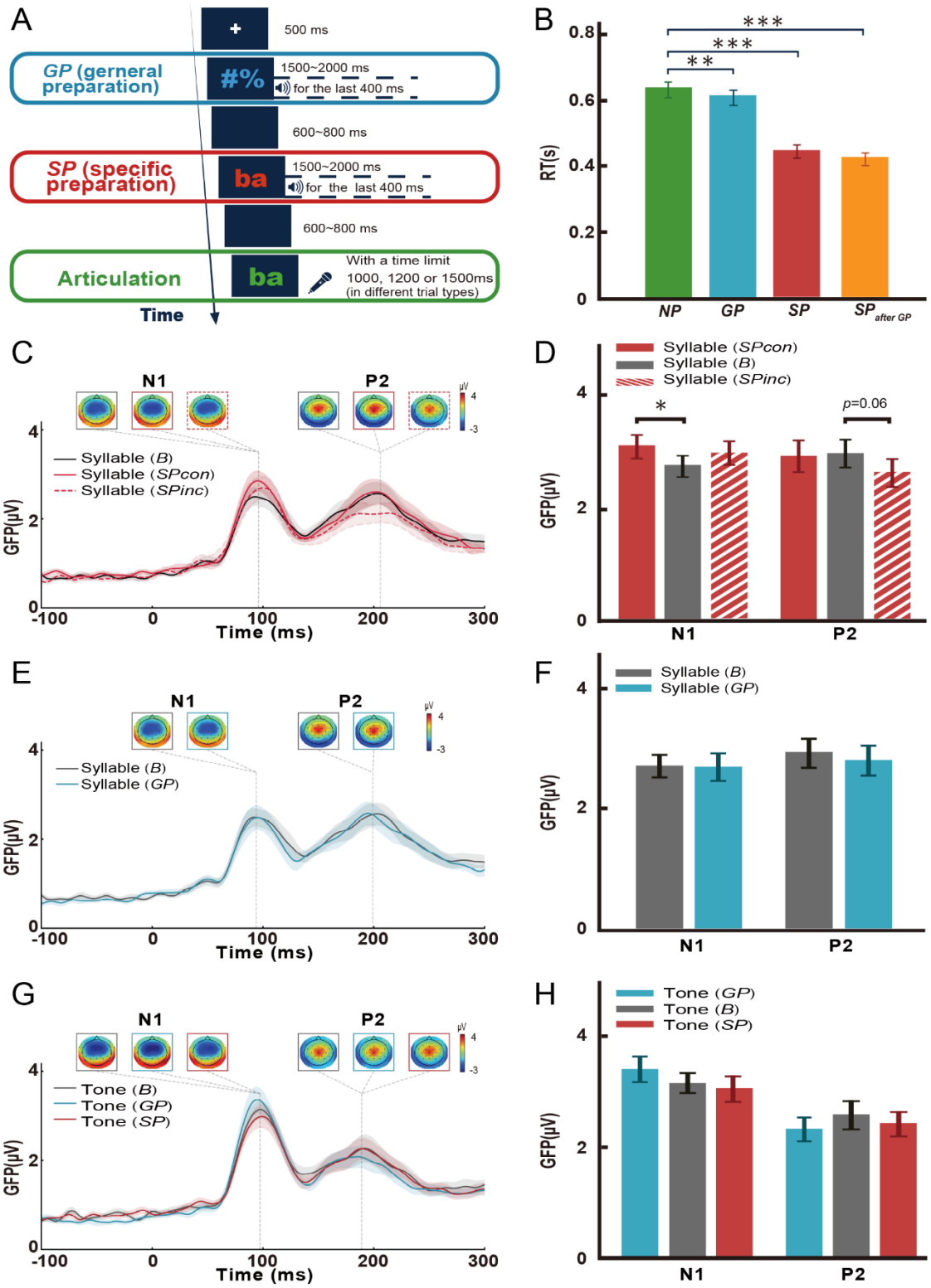

Figure 2. Experimental paradigm, behavioral, and ERP results for Experiment 1.

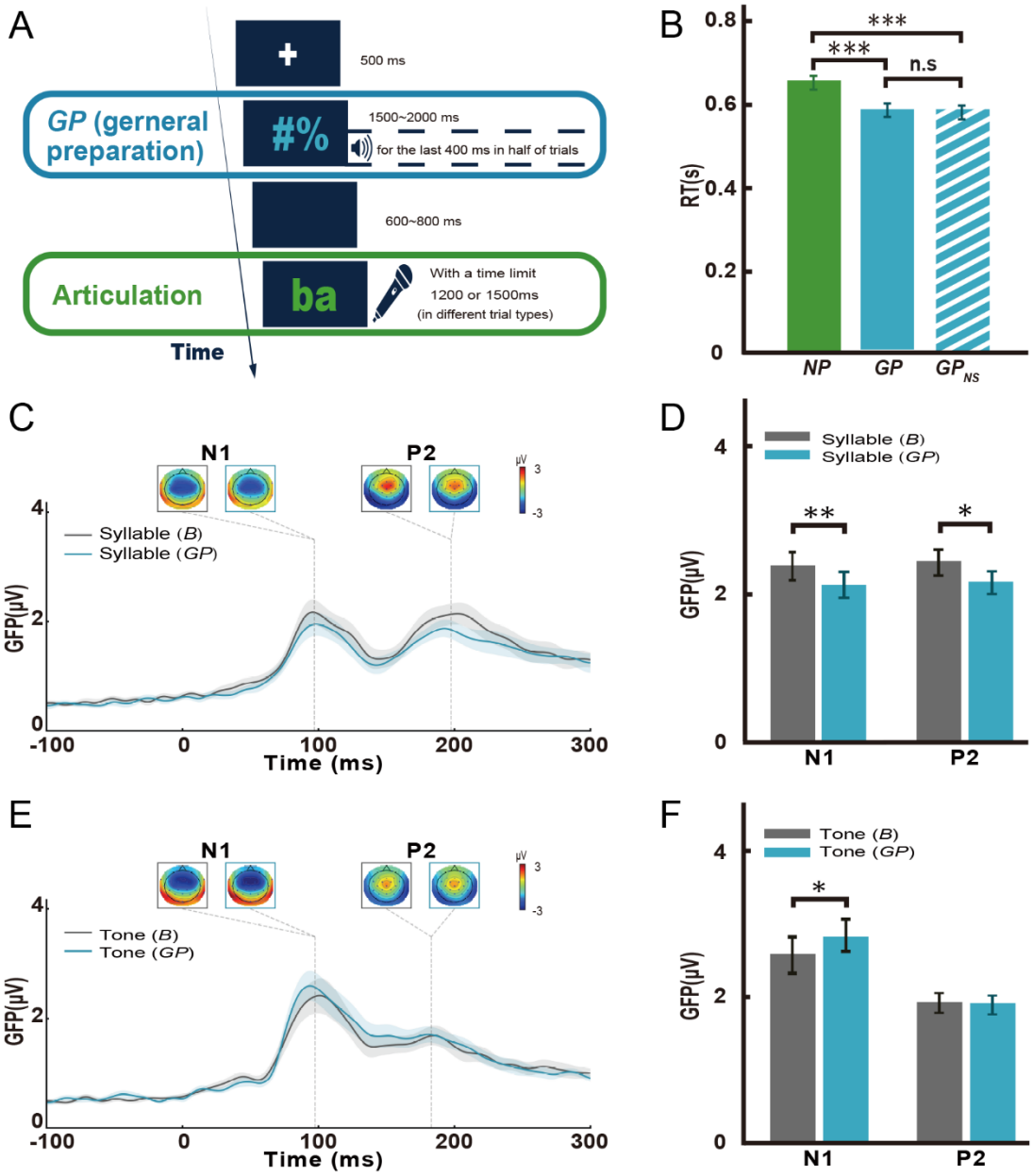

Figure 3. Experimental paradigm, behavioral, and ERP results for Experiment 2.

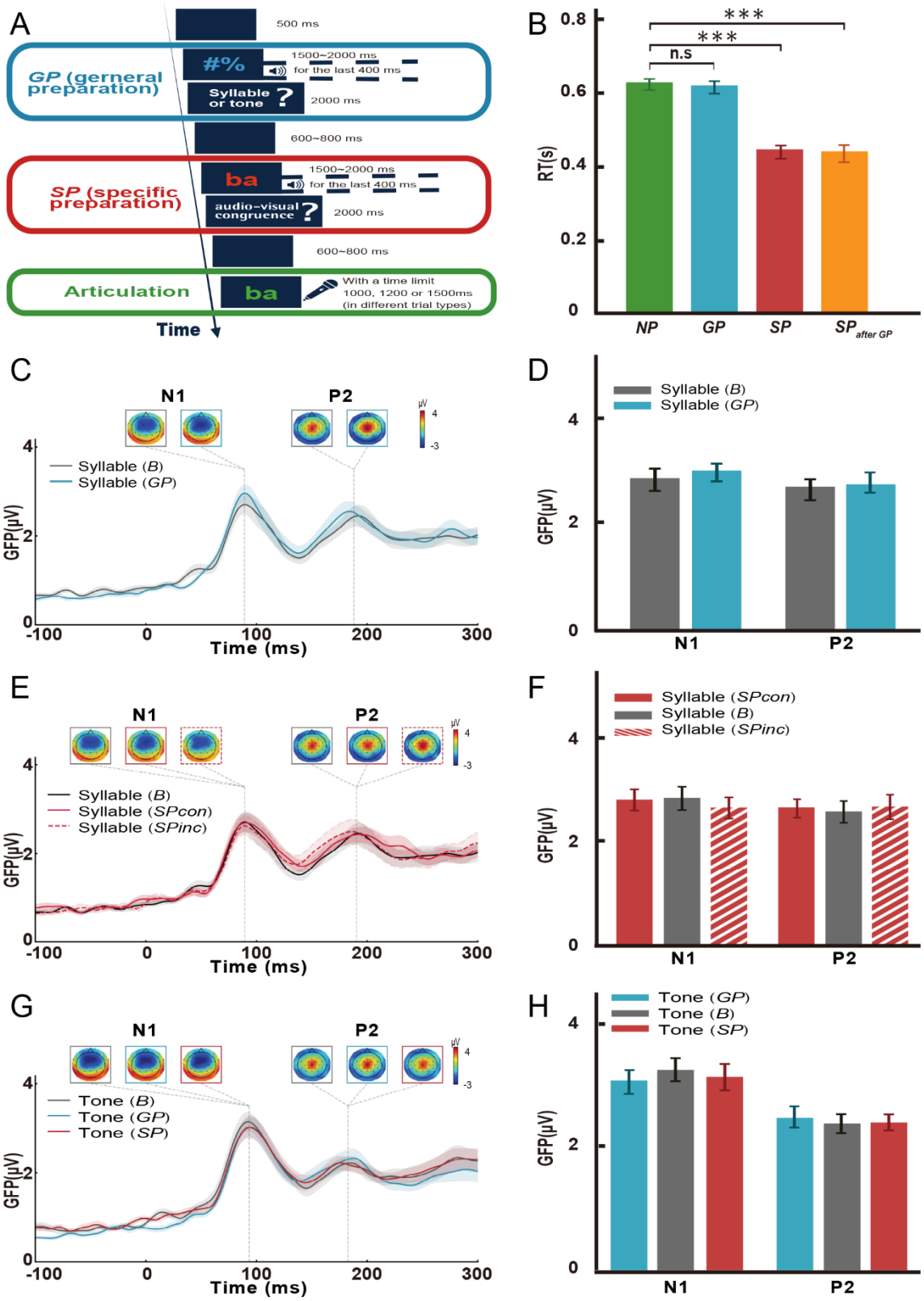

Figure 4. Experimental paradigm, behavioral, and ERP results for Experiment 3.

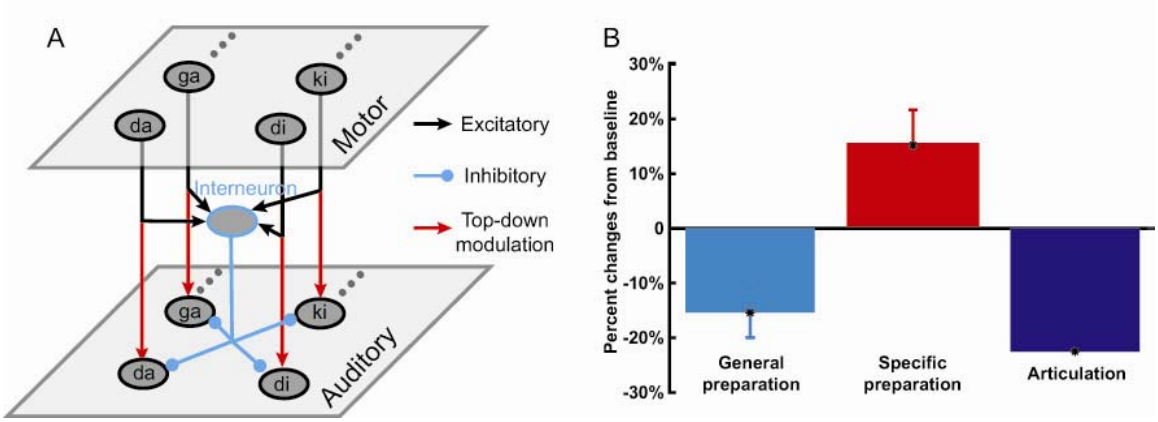

Figure 5. A neural network model of distinct motor signals and simulation results.