HHS Public Access

Author manuscript

J Cogn Neurosci. Author manuscript; available in PMC 2015 June 14.

Published in final edited form as:

J Cogn Neurosci. 2015 February ; 27(2): 352–364. doi:10.1162/jocn_a_00692.

Dynamics of Self-monitoring and Error Detection in Speech Production: Evidence from Mental Imagery and MEG

Xing Tian and **David Poeppel**

New York University

Abstract

A critical subroutine of self-monitoring during speech production is to detect any deviance between expected and actual auditory feedback. Here we investigated the associated neural dynamics using MEG recording in mental-imagery-of-speech paradigms. Participants covertly articulated the vowel /a/; their own (individually recorded) speech was played back, with parametric manipulation using four levels of pitch shift, crossed with four levels of onset delay. A nonmonotonic function was observed in early auditory responses when the onset delay was shorter than 100 msec: Suppression was observed for normal playback, but enhancement for pitch-shifted playback; however, the magnitude of enhancement decreased at the largest level of pitch shift that was out of pitch range for normal conversion, as suggested in two behavioral experiments. No difference was observed among different types of playback when the onset delay was longer than 100 msec. These results suggest that the prediction suppresses the response to normal feedback, which mediates source monitoring. When auditory feedback does not match the prediction, an "error term" is generated, which underlies deviance detection. We argue that, based on the observed nonmonotonic function, a frequency window (addressing spectral difference) and a time window (constraining temporal difference) jointly regulate the comparison between prediction and feedback in speech.

INTRODUCTION

A competent perception system must distinguish self-produced from externally generated perceptual events. Moreover, the consequences associated with inaccurately executed actions must be identified. A common mechanism for these source monitoring and error detection functions has been proposed in the framework of internal forward models (Wolpert & Ghahramani, 2000). The perceptual consequences of planned motor commands are predicted internally and compared with the perceptual feedback generated by the overt actions. Specifically, if prediction matches feedback, the neural responses to the external feedback are "canceled" and the perceptual changes are classified as self-produced. Evidence supporting such a cancellation mechanism has been found in visual (Sommer & Wurtz, 2002, 2006, 2008), tactile (Blakemore, Wolpert, & Frith, 1998, 2000), and auditory (Chang, Niziolek, Knight, Nagarajan, & Houde, 2013; Ventura, Nagarajan, & Houde, 2009; Eliades & Wang, 2003, 2005; Houde, Nagarajan, Sekihara, & Merzenich, 2002; Numminen,

Reprint requests should be sent to Xing Tian, 6 Washington Pl. STE 275, New York, NY 10003, or via xing.tian@nyu.edu..

Salmelin, & Hari, 1999) domains. Moreover, when internal prediction does not match feedback, greater auditory responses to perturbed online feedback are observed in speech production studies (Chang et al., 2013; Greenlee et al., 2013; Behroozmand, Liu, & Larson, 2011; Eliades & Wang, 2008; Tourville, Reilly, & Guenther, 2008), representing the discrepancy between the internal auditory prediction and overt feedback (neural error term).

The comparison between internal prediction and overt feedback requires characterization. For example, focusing on speech, does the increase in response magnitude to perturbed feedback (e.g., along the spectral dimension) correspond linearly to the degree of mismatch between prediction and feedback? Hints of nonlinearity come from electrophysiological studies using a pitch shift manipulation. The linear increase of neural responses as a function of pitch perturbation level stopped at a certain point (e.g., plateaued between 200 and 500 cents, Liu, Meshman, Behroozmand, & Larson, 2011; 250 cents in P2 responses, Scheerer, Behich, Liu, & Jones, 2013). Therefore, we hypothesize that a spectral integration window constrains the comparison between internal prediction and external feedback; hence, the neural responses to perturbed feedback do not linearly increase with spectral distance between prediction and feedback beyond the limit of the spectral window.

Moreover, how is the timing offset between prediction and feedback handled? The concept of temporal integration windows has been proposed for speech perception (Hickok & Poeppel, 2007; Poeppel, 2003). That is, information in a certain time range is chunked and integrated to form coherent representations. Audiovisual multisensory studies also suggest time constants of up to ~200 msec (van Wassenhove, Grant, & Poeppel, 2007; Munhall, Gribble, Sacco, & Ward, 1996). Another motivation for proposing the temporal integration windows comes from the phenomenon of delayed auditory feedback (DAF; Fairbanks, 1955; Black, 1951). Performance is deteriorated most when a delay of about 200 msec is introduced between speaking and hearing the self-produced sounds (e.g., Howell & Archer, 1984). The notion of a temporal integration window in DAF seems to underlie the integrity of perception and production processes. We build on this notion and hypothesize that a restrictive temporal integration window exists for the comparison between top–down (prediction) and bottom–up (feedback) representations. Specifically, the prediction about perceptual consequences of actions compares to the actual feedback only if the time lag between them is within the range of the temporal integration window. However, if the temporal distance is beyond the limit of such a window, the probability of feedback being self-produced will be very low; hence, no comparison will be carried out.

To link these hypotheses, we argue that a control mechanism (cf. Houde & Nagarajan, 2011; Grush, 2004), termed Kalman gain (K) from optimal control (Kalman, 1960), is applied over the neural error term (e_f; Figure 1). That is, taking an analog of the Kalman filter approach, by which more precise estimates of an unknown variable can be produced by using a series of measurements observed over time, the current observation of neural errors can be constrained by the stored distributions of the feedback's (temporal and spectral) characteristics. Specifically, we model this control mechanism as a function of spectral (D) and temporal (T) distances between internal prediction and external feedback (Equation 1):

$$K(T,D)=\begin{cases}\text{if} \quad T \geq \text{time}_{\text{threshold}} \rightarrow 0 \\ \text{if} \quad T<\text{time}_{\text{threshold}} \rightarrow \begin{cases} C & \text{if} \quad D<freq_{\text{threshold}} \\ f(D) & \text{if} \quad D \geq freq_{\text{threshold}} \end{cases}\end{cases} \quad (1)$$

If the time difference between the onset of prediction and feedback (*T*) is beyond the limit of a temporal integration window ($\text{time}_{\text{threshold}}$), no spectral comparison will be carried out, presumably because feedback with delays that are beyond the temporal threshold will not be perceived as self-produced, and therefore, the spectral properties of these sounds would not be relevant for the talker. If *T* is within $\text{time}_{\text{threshold}}$, the comparison will be carried out along the spectral dimension; if the spectral distance (*D*) is within the limit of a spectral integration window ($\text{freq}_{\text{threshold}}$), the gain function is a positive constant (*c*), and therefore, the error (e_f) caused by spectral discrepancy will be a linear function of *D*; if the spectral distance is beyond the limit, the gain will be a decreasing function of spectral distance, *f*(*D*), which scales down the error. Our simple model agrees with the core functions for self-monitoring and speech control in other computational models (e.g., Hickok, 2012; Houde & Nagarajan, 2011; Guenther, Ghosh, & Tourville, 2006; Grush, 2004). The current study is a new experimental approach aiming to provide direct neural evidence for the suggested computational mechanisms. Therefore, we test this model in behavioral experiments and an MEG study. Our aim is to determine the proposed values for temporal and spectral integration windows.

METHODS

Behavioral Assessment of Normal Voice Pitch Range

Normal conversation is the most common situation in which speech articulation occurs. Therefore, we assessed the spectral integration window using the subjective evaluation of voice pitch range in normal conversation. We used two behavioral measures to quantify the subjective probability of different levels of pitch shift occurring in one's normal pitch range, which provided the psychophysical evidence for the frequency boundary of spectral integration window.

Behavioral Experiment 1. Probability Judgment of Pitches Being in Normal Speech

Participants: Sixteen volunteers (eight men, mean age = 28.1 years, range = 24–47 years) took part in this experiment. All participants were right-handed. This experiment was approved by the New York University institutional review board.

Materials: Auditory stimuli were recorded in a quiet room using Radio Shack unidirectional dynamic microphone 33-3002. Participants pronounced the vowel /a/ 10 times using their normal, most comfortable pitch. The continuous auditory signals were recorded (sampling rate of 44.1 kHz) and further processed using Praat. Participants wore Sennheiser HD280 headphones when listening to the continuous recording and selected one auditory token as stimulus. The mean duration of the selected auditory tokens was 340 msec. All sounds were normalized by average intensity (RMS) to 70 dB SPL. Pitch shifts were introduced by using

the built-in pitch-shifting function in Praat (PSOLA). Specifically, the pitch was extracted using an auto-correlation method (Boersma, 1993) and shifted to a given level, and the shifted pitch contour was used to resynthesize the vowel sound using the overlay-and-add method (Moulines & Charpentier, 1990). Specifically, nine levels of pitch manipulation were applied to the original recordings, resulting in nine conditions with −30, 150, 300, 450, 600, 750, 900, 1050, and 1200 cent pitch shifts. The cent is a logarithmic unit of measure for relative frequency difference, which is computed as Cents = $1200 \times \log_2(F_T/F_B)$, where F_T is the target frequency (in Hz) and F_B is the baseline frequency (the individual recorded normal pitch in our case). One hundred cents equals one semitone, and 1200 cents is an octave. Because of the comb-filtering effect caused by the timing difference between bone conduction and air conduction, hearing one's own voice during speech production has a lower perceived frequency than the perception of a recorded voice (Shearer, 1978). Therefore, the pitch of the recorded sounds was shifted down by 30 cents to make them more similar to their own voice (Shuster & Durrant, 2003). All sounds were normalized and delivered at about 70 dB SPL.

Procedure: One hundred forty-four trials (16 trials per pitch shift) were randomly presented in a block. Participants were asked to judge the probability of a given sound being in their normal conversational voice range on a 9-point scale. Specifically, the response “9” stood for the most common pitch during everyday conversation (Probability 1), whereas the response “1” represented the most impossible pitch (Probability 0), with the occurrence probability proportionally decreasing from 9 to 1. The midpoint “5” was highlighted to be the 50% point. That is, the occurrence of this particular pitch during normal conversation is at chance: Participants may or may not use that particular pitch to ask a question. A pitch that is higher than the pitch at the 50% point is less likely to occur during normal conversion and should be assigned to one of the responses in 1–4; a pitch that is lower than the pitch at the 50% point is more likely to occur during normal conversion and should be assigned to one of the responses in 6–9. The anticorrelation between number response keys and levels of pitch shift (“9” for the most common voice that has the lowest pitch and “1” for the most unlikely voice that presumably has the highest pitch) was used to avoid the explicit association of pitch shift direction and number scale.

Data analysis: Probability responses were recorded and linearly transformed to percentages (response “9” to 100% and response “1” to 0%, with linear decrement of 12.5% per step). The percentage scores were averaged across 16 trials for each level of pitch-shifted sound.

Behavioral Experiment 2. Categorical Judgment of Pitches Being in Normal Speech—Participants could implement strategies during the first behavioral experiment. For example, they could remember the nine levels of pitch shifting and assign the numeric responses to each of them, although postexperiment interviews revealed that most participants only distinguished and memorized at most five levels, suggesting that such a strategy was less likely. Moreover, the probability judgment for pitch may create a bias when making judgments of less probable events (although that is not our main measure of interest), because it is hard to judge the possibility of occurrence for a pitch that would be

out of one's voice range during normal conversation. Therefore, we ran a second behavioral experiment using a categorization task to assess the pitch range during normal conversation.

Participants and materials: The same 16 volunteers as in Behavioral Experiment 1 participated in this experiment.

Procedure: The experimental procedure was identical to Behavioral Experiment 1, except the responses were reduced to three categories. Specifically, participants were asked to make a judgment whether a given pitch was in their voice pitch range during normal conversation. They were required to press the button “9” for the pitches in their voice range, whereas the button “1” was pressed for pitches out of their vocal range. The button “5” was assigned to the pitches that participants were unsure of. That is, these particular pitches were on the edge of the highest pitch they would use during normal conversation (when asking a question). Any pitch that was higher than the ones assigned to the “5” response should be out of their normal voice range, and any pitch that were lower than the ones assigned to the “5” response should be in their normal voice range.

Data analysis: Probability responses were recorded and transformed to percentages (e.g., response “9” to 100% and response “5” to 50% and response “1” to 0%). The percentage scores were averaged across all 16 trials for each pitch shifted sound.

MEG Experiment. Pitch Shift and Playback Onset Delay during Articulation Imagery

We used articulation imagery (imagine speaking without moving any articulators or making any sound), which has been modeled as a prediction process (via the internal simulation and estimation mechanism) without overt output overlaps (Tian & Poeppel, 2010, 2012; Grush, 2004). We deploy this paradigm here to investigate the neural response to parametric variation in internal–external mismatch by introducing both pitch perturbation and onset delay. The interaction between articulation imagery and perception has been demonstrated in recent studies using behavioral measures (Berger & Ehrsson, 2013; Scott, 2013) and electrophysiological recordings (Tian & Poeppel, 2013). The proposed temporal and spectral integration windows in the control mechanism leads us to hypothesize (a) that auditory neural response magnitude is a nonmonotonic function of spectral distance between the internal prediction and pitch-shifted playback and (b) that such comparison is constrained by the temporal distance between the occurrence of prediction and playback.

Participants—Sixteen individuals (seven men; mean age = 27.5 years, range = 19–44 years) participated in this experiment for monetary compensation. All participants were right-handed and without history of neurological disorders. This experiment was approved by the New York University institutional review board.

Materials—Same recording procedures were used as in Behavioral Experiment 1. Four different levels of pitch shift were selected and used in this experiment. These four levels were −30, 300, 600, and 1200 cent pitch-shifted sounds. We chose only upward shifts to save experimental time in the electrophysiological recordings. Results from previous studies suggest that human electrophysiological responses to downward and upward shifts have no

qualitative (only quantitative) differences (e.g., upward shifts: Behroozmand, Karvelis, Liu, & Larson, 2009; downward shifts: Scheerer et al., 2013). Considering our experimental design that includes an additional factor of onset delay with four levels (see Procedure below), we only chose the upward shifts to minimize the total number of conditions and reduce the MEG recording time. Second, our aim was to investigate the existence of integration windows for self-monitoring in speech. The quantitative differences found between downward and upward shifts demonstrate that people would be very sensitive to the pitch decreasing. That is, the distribution of downward shifts for the hypothetical spectral integration window would be very narrow. Therefore, we chose upward shifts to increase sensitivity for testing the spectral integration window.

Procedure—Participants were asked to imagine articulating /a/ using their normal and most comfortable pitch, same as during recording before the experiment. They were instructed to press a button using their left index finger to indicate the beginning of articulation imagery in each trial (Figure 2). The button press triggered one of the four pitch shifted sounds with one of the four levels of onset delay ([0, 100, 200, 500] msec). Thus, 16 conditions were run (four levels of pitch shift crossed with four levels of onset delay). Participants were asked to passively listen to the playback and to start the next trial after the offset of playback sound, at a comfortable self-paced speed. This procedure was used to prevent overlap between the auditory responses to the playback and motor responses at the beginning of the next trial. Participants were familiarized with the procedure by training before the experiment. They were also trained on the task of articulation imagery, and all participants confirmed they could induce the quasi-kinesthetic and auditory experiences vividly, without physically moving any articulators. We set a microphone next to participants to monitor that there was no overt pronunciation throughout the experiment. The observations of overlapping neural networks between covert and overt movement in motor imagery studies (e.g., Dechent, Merboldt, & Frahm, 2004; Meister et al., 2004; Ehrsson, Geyer, & Naito, 2003; Hanakawa et al., 2003; Gerardin et al., 2000; Lotze et al., 1999; Deiber et al., 1998) support that both types of articulator movement induce a similar motor efference copy, as suggested in numerous theoretical pieces (e.g., Desmurget & Sirigu, 2009; Grush, 2004; Miall & Wolpert, 1996; Jeannerod, 1994, 1995). As long as there is no overt sound, our goal of an internally induced auditory representation from a motor efference copy is valid. Potential subvocal movement is irrelevant to the interpretation.

Twenty-five blocks were included in the experiment, with 32 trials in each block (2 trials per condition in each block, 50 trials per condition in total). The presentation order was randomized. Each block began with three 250-Hz sinusoidal tones followed by three original voice stimuli (−30 cents pitch shifted) to remind participants of their normal voices and the voice they should imagine during the task of articulation imagery.

Twenty-five blocks of a baseline task, with same trial number, were also run. The baseline section used the identical procedure as the main experiment, except participants did not perform articulation imagery. Participants pressed a button and passively listened to the playback. They were instructed to proceed at their comfortable speed and only to press the button after the offset of the playback in the last trial. The order of the main experiment and the baseline task was counterbalanced across participants. Because the baseline run and

main experiment were balanced in terms of auditory memory, motor temporal prediction, and other nuisance variables, only the interaction between prediction and auditory process was available in the main experiment. Therefore, comparison between auditory responses in main experiment and baseline run will point to the mechanism of internal–external interaction.

Three additional control runs were carried out. In one run (auditory control), participants passively listened to the four different pitch shifted sounds 50 times each in a random order. This auditory control was to test whether the magnitude of auditory responses to equal-loudness but different-pitch sounds was similar and hence to rule out that any differences between main experiment and baseline run were because of pitch differences in auditory playback. In the other two runs (motor control), participants were asked to press a button with articulation imagery in one run, whereas they only pressed a button in another run. These two motor control runs included 50 trials each and were compared with each other to test whether the magnitude of motor responses that overlapped with the auditory responses during the experiment was independent from task demands and hence rule out that any differences between main experiment and baseline run were because of motor and imagery interaction. All the control runs were carried out after the main experiment and baseline run. The auditory control was run first, which also served as a break. Then the two motor control runs were carried out and counterbalanced across participants.

MEG Recording—Neuromagnetic signals were measured using a 157-channel whole-head axial gradiometer system (KIT, Kanazawa, Japan). Five electromagnetic coils were attached to a participant's head to monitor head position during MEG recording. The locations of the coils were determined with respect to three anatomical landmarks (nasion, left and right preauricular points) on the scalp using 3-D digitizer software (Source Signal Imaging, Inc., La Mesa, CA) and digitizing hardware (Polhemus, Inc., Colchester, VT). The coils were localized to the MEG sensors at both the beginning and the end of the experiment. The MEG data were acquired with a sampling frequency of 1000 Hz, filtered online between 1 and 200 Hz, with a notch filter at 60 Hz.

MEG Analysis—Raw data were noise-reduced offline using the time-shifted PCA method (de Cheveigné & Simon, 2007). Trials with amplitudes of >2 pT (~5%) were considered artifacts and discarded. For each condition, epochs of response to the auditory playback, 600 msec in duration including a 100-msec prestimulus period, were extracted and averaged in both articulation imagery and baseline runs. The averages were low-pass filtered with a cutoff frequency of 30 Hz. The typical M100/200 auditory response complex was observed (Roberts, Ferrari, Stufflebeam, & Poeppel, 2000), and the peak latencies were identified for each individual participant.

Because of possible confounds between the changes in neural source magnitude and changes in neural sources distribution during analyses at the sensor level (Tian & Huber, 2008), a multivariate measurement technique (angle test of response similarity), developed by Tian and Huber (2008) and recently available as an open-source toolbox (Tian, Poeppel, & Huber, 2011), was implemented to assess the topographic similarity between auditory responses to the playback in experimental (articulation imagery) and baseline (without

articulation imagery) runs. This technique allows the assessment of spatial similarity in electrophysiological studies regardless of response magnitude and estimates the similarities in underlying neural sources distribution (e.g., Luo, Tian, Song, Zhou, & Poeppel, 2013; Tian & Huber, 2013; Tian & Poeppel, 2010, 2013; Davelaar, Tian, Weidemann, & Huber, 2011; Huber, Tian, Curran, O'Reilly, & Woroch, 2008). In this method, each topographical pattern is considered as a high-dimensional vector, where the number of dimensions equals the number of sensors in recording. The angle between the two vectors represents the degree of similarity/difference between the two topographies. The cosine value of this angle, which is called *angle measure*, can be calculated from the dot product of these two response vectors where the value "1" stands for exact match (angle equals zero) and the value "−1" stands for opposite (angle equals π).

The angle measure between topographies in different conditions, termed *between angle measure*, is statistically tested against a null hypothesis (i.e., is the angle between two topographic patterns greater than chance). The null hypothesis is formed by comparing the pattern similarity of average responses for the first half and second half of the experiment within each condition (within angle measure, which is the maximum similarity value after taking the systematic noise, such as fatigue and movement into account). In this study, the between angle measure was calculated between auditory responses in the imagery run versus in baseline for each of 16 conditions. The between angle measure was compared with the within angle measure of that condition to statistically determine the topographic similarity between auditory responses in experimental and baseline runs for each condition. If the between angle measure is significantly smaller than the within angle measure (the angle between two topographies is larger than chance), the two topographies are different and hence infer distinct neural sources distribution, whereas if they are not significantly different, the null results suggest the two topographies are similar and the following magnitude test could be free of confounds of source distribution changes. Although the null results could be caused by lack of power, this issue can be ruled out by significant results in the following statistical tests that have the same power.

After confirming the stability of the neural source distributions of auditory responses across experimental and baseline runs, any observed significant effects in sensor level analysis will contribute to the response magnitude change. The root mean square (RMS) of waveforms across 157 channels, indicating the global response power in each condition, was calculated and employed in the following statistical tests. A 25-msec time window centered at individual M100 and M200 latencies was applied to obtain the temporal average responses. The relative response power changes were further calculated by subtracting the responses to auditory playback during the imagery run from the one during the baseline run for each condition. A repeated-measures two-way ANOVA was carried out on the factors Pitch shift and Onset delay. Four separate repeated-measures one-way ANOVAs were further implemented on the factor of Pitch shift in each of four levels of Onset delay. Planned one-sample *t* tests (one-tailed) were carried out for each level of pitch shift, and planned comparison paired *t* tests were carried out between two adjacent levels of pitch shift. The employment of the one-tailed test in each condition followed directly from our specific hypothesis. Perceptual suppression has been found in the responses to the normal feedback

during actions (e.g., Chang et al., 2013; Christoffels, van de Ven, Waldorp, Formisano, & Schiller, 2011; Houde et al., 2002; Numminen et al., 1999; Blakemore et al., 1998), whereas enhancement has been found whenever the feedback deviates from actions (Chang et al., 2013; Scheerer et al., 2013; Behroozmand et al., 2009; Tourville et al., 2008). Therefore, we hypothesized a specific modulation direction for a particular condition: Suppression when the playback is normal, and enhancement when the playback is altered.

To test the consistency of motor response magnitude under different task demands (with or without articulation imagery), a 25-msec time window centered at the button press and release responses was applied to obtain the temporal average of motor responses in different tasks. Paired *t* tests were run on the average responses between runs with and without articulation imagery. To test the magnitude consistency of auditory responses to different levels of pitch shifted sounds, a 25-msec time window centered at early M100 and M200 auditory responses in the passive listening run was applied to obtain the temporal average of auditory responses to different sounds. A repeated-measures one-way ANOVA was carried out on the factors of Pitch shift for both M100 and M200 components.

RESULTS

In the behavioral experiments, a repeated-measures one-way ANOVA revealed that the probability of a pitch being in one's own voice range during normal conversation is significantly different across the different levels of pitch shift, in both Behavioral Experiment 1 using probability judgment [$F(8, 120) = 88.34$, $p < .001$] and in Behavioral Experiment 2 using categorization judgment [$F(8, 120) = 131.08$, $p < .001$] (Figure 3). Specifically, the rating of a given pitch as being in one's voice range decreases as the degree of pitch shift increases, indicated by the significant linear contrast [$F(1, 15) = 138.14$, $p < .001$ in Experiment 1; $F(1, 15) = 322.92$, $p < .001$ in Experiment 2]. Moreover, the quadratic components were also significant [$F(1, 15) = 10.81$, $p < .01$ in Experiment 1; $F(1, 15) = 14.74$, $p < .005$ in Experiment 2], suggesting the fast rate of decrease at the larger degree of pitch shift. The cubic component was significant in Behavioral Experiment 2 [$F(1, 15) = 5.56$, $p < .05$], suggesting fewer possible response categories made the separate criterion of “within” and “out of” normal voice range clearer and the transition between the rating of in and out pitch range sharper. More importantly, the one-sample *t* tests against the chance level (0.5) revealed that the 600-cent pitch shift was rated as in one's normal pitch range [$t(15) = 3.97$, $p < .005$ in Experiment 1; $t(15) = 3.71$, $p < .005$ in Experiment 2], whereas the 1200-cent pitch shift was clearly out of range [$t(15) = -4.27$, $p < .001$ in Experiment 1; $t(15) = -13.93$, $p < .001$ in Experiment 2]. For other levels between 600 and 1200 cents, the 750-cent pitch shift was rated not significantly different from the chance level [$t(15) = 1.62$, $p > .13$ in Experiment 1; $t(15) = 0.34$, $p > .74$ in Experiment 2]; the 900-cent level was at chance level in Experiment 1 [$t(15) = -0.45$, $p > .66$], but was rated out of range in Experiment 2 [$t(15) = -2.27$, $p < .05$]; the 1050-cent level was out of range in both experiments [$t(15) = -2.50$, $p < .05$ in Experiment 1; $t(15) = -5.98$, $p < .001$ in Experiment 2]. Therefore, the pitch boundary in subjective probability judgment of one's voice range in normal conversation is higher than 600 cents and lower than 1200 cents, consistent with the proposed spectral integration window.

In the MEG experiment, the RMS waveforms were calculated using all channels for each condition during experimental and baseline runs. As shown in Figure 4, early auditory response patterns (M100/M200 complex) were observed in all conditions. These waveforms exhibited trends in the response magnitude changes associated with levels of pitch manipulation within each level of onset delay. Specifically, in the immediate playback condition (0-msec delay), normal playback elicited smaller responses during the imagery task, whereas the auditory response was enhanced in all the pitch-shifted conditions. A similar trend was observed in the 100-msec onset delay conditions. However, for the two other onset delay levels (200 and 500 msec), no suppression effects were observed.

Before statistically testing the magnitude changes between the imagery and baseline runs, the similarity between topographies that reflect the underlying neural source distributions were quantified. The auditory response patterns of the M100 and M200 components were calculated for auditory playback in all 16 conditions during experimental and baseline runs (see Figure 5 for M100; the pattern for the M200 is not shown). Importantly, the angle test did not reveal any significant spatial pattern differences between responses to auditory playback in the imagery and baseline runs for any condition (*ps* > .05). That is, the topographies of auditory responses were highly similar in both runs and the distribution of neural sources that mediated auditory perception was independent from task demands. Thus, the following magnitude tests will be free from changes in neural source distribution.

The effects of prediction during imagery on playback were quantified by the time averaged data around individual M100 and M200 peak latencies. For the M100 component, clear interaction patterns were observed (Figure 6). A repeated-measures two-way ANOVA on the factors of Pitch shift and Onset delay revealed that the main effect of Pitch shift was significant [$F(3, 45) = 6.29$, $p < .005$] and the interaction between Pitch shift and Onset delay was also significant [$F(3, 45) = 2.96$, $p < .005$]. The trend of response changes caused by the factor Pitch shift was further tested within each level of onset delay using a repeated-measures one-way ANOVA. The main effects were significant at the onset delay levels of 0 msec [$F(3, 45) = 11.38$, $p < .001$] and 100 msec [$F(3, 45) = 8.18$, $p < .001$], but not at the levels of 200 msec or 500 msec (*F*s < 1). Therefore, further tests were run only at the onset delay levels of 0 and 100 msec. Planned one sample *t* tests (one-tailed) revealed that the prediction-induced suppression occurred for normal playback in the onset delay level of 0 msec [$t(15) = -2.08$, $p < .05$] and was marginal at 100 msec [$t(15) = -1.56$, $p = .07$]; whereas when the playback was pitch shifted, enhancement were observed in the onset delay levels of 0 msec [for 300-cent pitch shift, $t(15) = 1.955$, $p < .05$; for 600-cent pitch shift, $t(15) = 4.49$, $p < .001$; and for 1200-cent pitch shift, $t(15) = 2.46$, $p < .05$] and 100 msec [for 300-cent pitch shift, $t(15) = 2.07$, $p < .05$; for 600-cent pitch shift, $t(15) = 4.98$, $p < .001$; and for 1200-cent pitch shift $t(15) = 2.36$, $p < .05$].

The linear and quadratic contrasts were also significant at the onset delay levels of 0 msec [$F(1, 15) = 14.44$, $p < .005$; $F(1, 15) = 25.41$, $p < .001$] and 100 msec [$F(1, 15) = 11.67$, $p < .005$; $F(1, 15) = 11.70$, $p < .005$]. Planned comparison paired *t* tests revealed that auditory responses to normal playback were significantly smaller than the ones to 300-cent pitch-shifted playback at the onset delay levels of 0 msec [$t(15) = -3.70$, $p < .005$] and 100 msec [$t(15) = -2.89$, $p < .01$]; responses to 300-cent pitch-shifted playback were also significantly

smaller than the ones to 600-cent playback at the onset delay levels of 0 msec [$t(15) = -2.64$, $p < .01$] and 100 msec [$t(15) = -2.22$, $p < .05$]. However, a significant trend of response amplitude decrease from responses to 600 cents to 1200 cents pitch shift was observed at the onset delay levels of 0 msec [$t(15) = 2.08$, $p < .05$] and 100 msec [$t(15) = 1.77$, $p < .05$]. These results suggest that the auditory M100 response to pitch-shifted playback during articulation imagery is a nonmonotonic function of the frequency distance between prediction and playback: The response magnitude linearly increases as the frequency distance increases, but decreases at the largest distances. This transition of auditory responses to pitch-shifted playback between 600 and 1200 cents is consistent with the pitch boundary obtained in behavioral experiments. Furthermore, the temporal distance between prediction and playback constrains the changes of M100 response magnitude as a function of frequency differences, indicated by the absence of variation beyond the onset delay of 100 msec. Therefore, the observed nonmonotonic functions both in time and frequency dimensions were consistent with proposed temporal and spectral integration windows in a gain control mechanism for the neural error term that reflected the differences between prediction and playback.

The effects of pitch shift and onset delay did not extend to the M200 response (Figure 6). A repeated-measures two-way ANOVA on the factors of Pitch shift and Onset delay did not reveal any significant main effects or interaction (Fs < 1). The following repeated-measures one-way ANOVA at all onset delay levels did not reveal any significant effects caused by the factor of Pitch shift (Fs < 1). Neither was the linear contrast (ps > .30) nor quadratic contrast significant (Fs < 1).

A repeated-measures one-way ANOVA on the passive listening control run suggested that the magnitude of M100 and M200 auditory responses was not different across the four types of sounds with different pitches (Fs < 1; Figure 7). Moreover, a paired t test between button press responses accompanied with or without articulation imagery did not reveal any difference [for muscle extraction component (button press), $t(15) = 0.33$, $p > .75$; for muscle relaxation component (button release), $t(15) = -0.38$, $p > .71$] (Figure 7), suggesting that the neural responses associated with button press that overlapped with early auditory responses at 0-msec time delay conditions were independent from task demands, and any differences observed in those conditions were caused by the prediction in imagery task and manipulation of pitch shift in playback.

DISCUSSION

This study investigates the comparison mechanism between internal prediction and external playback in (covert) speech production. We used articulation imagery as a model and implemented a parametric manipulation of both pitch and onset delay playback. We observed that when overt playback was delayed by 200 msec or more relative to the initiation of articulation, the auditory responses to different pitch manipulations were not different. In contrast, when the delay was less than 200 msec, the responses were (i) suppressed when “normal” playback was provided or (ii) increased linearly as the degree of pitch shifted up to 600 cents.

This study, using covert imagined articulation, complements the findings from the commonly used overt speech feedback paradigms and adds a new dimension to the issue. Our results on (covert) articulation imagery-induced suppression and enhancement are consistent with the findings of overt production-induced suppression for normal feedback (Chang et al., 2013; Ventura et al., 2009; Eliades & Wang, 2003, 2005; Houde et al., 2002; Numminen et al., 1999) and enhancement for perturbed feedback (Chang et al., 2013; Greenlee et al., 2013; Behroozmand et al., 2011; Eliades & Wang, 2008; Tourville et al., 2008). This consistency demonstrates mental imagery as a valid method in research on action-perception interaction and its cognitive neural mechanisms. The advantage of mental imagery paradigms, such as elimination of overlaps in neural responses from overt action and feedback, comes at a cost. For example, the trial-by-trial variability is hard, if not impossible, to quantify. However, the measureable effects obtained in this study demonstrate the reliability of using mental imagery as a paradigm in human electrophysiological studies if the experiments are carefully designed and controlled. The implementation of articulation imagery as a model of auditory prediction in this study eliminates overt output and provides direct evidence supporting the cancellation by auditory prediction of normal auditory feedback as well as response increases indicating the discrepancy between manipulated feedback and prediction.

The pitch of the overt playback was manipulated in this study, and the effects of comparison between prediction and playback were found in the M100 but not the M200 response components. In a previous study, the phonological content (syllables) of playback was either congruent or incongruent with the preceding auditory prediction and there the manipulation only affected the M200 component but not the M100 (Tian & Poeppel, 2013). Therefore, the occurrence and attributes of the internal–external interaction in the auditory processing hierarchy depend on the context and level of the manipulation.

Another reason why the effects were absent in M200 component may be because of the lack of speech production task demands during mental imagery. For example, a recent EEG study showed that the P2 response (the analog of M200) correlated with vocal compensation response magnitude (Scheerer et al., 2013). That is, P2 could reflect the start of the sensory-to-motor transformation (update the motor plan to compensate the pitch shift), a downstream process from the N1 that reflects the comparison between prediction and playback. However, there is no sensory–motor transformation task demand in our imagery study, which could lead to the absent of M200 effect. The simplification of the task demands lets us focus on testing the comparison mechanism between efference and playback, which is another strength of the new imagery paradigm.

The nonmonotonic profile of the auditory response as the pitch shift increased during articulation imagery suggests that the comparison between prediction and playback is constrained by a spectral integration window and the error term, manifested in the response increase, is modulated by a control mechanism that is a function of spectral distance between the internal estimate and external stimulus. Specifically, if the prediction matches the playback, the responses to the external stimuli are suppressed, presumably because of cancellation of playback caused by the similar representation. As the deviance between prediction and playback increases, the auditory response to the playback increases till the

system detects that the difference is too large to be plausibly self-generated. The proposed limit of spectral integration window (spectral threshold as in Figure 1 and Equation 1) was in the range between 600 and 1200 cents, which correlated with the subjective assessment of pitch range in normal conversation determined in the behavioral experiments and the turning point in nonmonotonic function observed in MEG. These results suggest that if the playback is so different from the prediction that it is beyond one's normal range, the playback will be no longer treated as self-produced and the error terms will be scaled down. These results also agree with the observation of less compensation to larger F1 format shift (Katseff, Houde, & Johnson, 2012), consistent with the proposed spectral integration window and gain control function.

Our finding of a threshold at about 600 cents in the spectral integration window is consistent with observations in other electrophysiological studies. For example, the N1 responses were increased from 0 to 100 cents but plateaued from 100 to 250 cents and increased again to the maximum shift of 400 cents (Scheerer et al., 2013). In another EEG study, 500 cents yielded similar N1 responses magnitude as 200 cents (Liu et al., 2011). All these studies support that the “turning point” for the N1 is beyond 400 cents. To our knowledge, previous speech feedback alteration EEG studies did not test pitch shifted by more than 500 cents in normal participants. Therefore, our study extends previous findings and suggests the spectral threshold for self-monitoring in speech is about 600 cents. (See discussion below about altered auditory feedback and stuttering for more consistent evidence about the “turning points” at around 600 cents from the clinical population.)

We suggest that a temporal integration window also constrains the comparison between prediction and playback, demonstrated by the absence of the nonmonotonic variation of auditory responses to pitch shifted playback beyond the onset delay of 200 msec. Therefore, the duration of the window in which the prediction and playback may integrate is about 100–200 msec, approximately consistent with the long temporal integration window of 200 msec during speech perception (Hickok & Poeppel, 2007; Poeppel, 2003), indicating a similar temporal integration window for combining information from bottom–up and top–down processes. The integration window may be a ubiquitous temporal constraint on the interaction between abstract representations. Moreover, the temporal integration window alone can serve as a way to test causality: Any perceptual changes falling into this time limit after speech production will be treated as self-produced and compared with prediction for self-monitoring and error detection in online control. The suppression in the response to playback indicates self-production and enhancement represents differences between prediction and playback indicated that the self-produced sound is not as planned. Also note that all the responses to the pitch shifts at 200 and 500 msec temporal delays remained elevated, which could reflect that the temporal asynchrony between the efference copy in covert action and auditory playback was beyond the threshold of the temporal integration window.

The combination of spectral and temporal integration windows can solve the problem that occurs when externally induced sounds (usually different from self-produced speech) occur within the temporal window of internal prediction, which could induce the same response increase as the inaccurate self-generated speech feedback. Previous studies suggest that

extremely manipulated feedback, such as using noise to substitute normal speech feedback, did not induce response magnitude increases as compared with the passive listening responses (Christoffels et al., 2011; Houde et al., 2002). Only the perturbation of feedback in a subset of features (e.g., pitch, F1, F2 formants) show increased responses (Behroozmand et al., 2011; Eliades & Wang, 2008; Tourville et al., 2008). It could be that, in the case of extreme differences between feedback and prediction, the gain control mechanism scales down the error term to the normal response level as in passive listening. However, should partial features overlap but other features diverge between feedback and prediction, the gain control mechanism will not scale down the error term completely but proportionally to the amount of overlap (such as a function of spectral difference in Equation 1, if the difference is beyond certain threshold). The observation of an increased response to 1200 cents pitch shifted playback could be the case of mild magnitude scaling down because partial features (such as the speech envelope) were still overlapped between playback and prediction. Therefore, the spectral and temporal integration windows together with the gain control mechanism provide a feasible solution to avoid the problem of identifying externally induced sounds as inaccurate self-generated feedback and prevent unnecessary motor correction.

A similar gain control mechanism with the spectral and temporal integration windows could underlie stuttering amelioration when the normal speech feedback is substituted with either DAF (e.g., Vaxes, 1963; Naylor, 1953) or frequency altered feedback (FAF; e.g., Kalinowski, Armson, Stuart, & Gracco, 1993; Howell, El-Yaniv, & Powell, 1987): The scaled down error term, a consequence of the differences between internal prediction and external feedback, initiates less motor correction compared with the larger error term caused by normal speech feedback and noisy prediction in people who stutter (Tian & Poeppel, 2012; Hickok, Houde, & Rong, 2011; Max, Guenther, Gracco, Ghosh, & Wallace, 2004). Interestingly, a similar nonlinearity of fluency enhancement was observed in DAF and FAF. For DAF, 50- and 75-msec delays were found to enhance speech fluency to the same degree, as both of them produce better amelioration than 25-msec delay (Kalinowski, Stuart, Sark, & Armson, 1996). A similar enhancement plateau has been found for delays of 50–150 msec (Burke, 1975; Webster, Schumacher, & Lubker, 1970). For FAF, the fluency benefit trend linearly increases from zero to a half octave (600 cents) pitch shift (Antipova, Purdy, Blakeley, & Williams, 2008); but half an octave and one octave (1200 cents) FAF provide similar amelioration results (Hargrave, Kalinowski, Stuart, Armson, & Jones, 1994; also see Stuart, Kalinowski, Armson, Stenstrom, & Jones, 1996). This nonlinearity of fluency enhancement across levels of DAF and FAF suggests that the increasing discrepancy between prediction and feedback, both in temporal and frequency domains, increases the probability for the gain control function to scale down the errors. Yet beyond a certain degree of divergence, all levels of perturbation in feedback are treated equally extreme; hence, the scaling down factors hit a plateau, leading to similar fluency enhancement.

Using a unique pairing of a mental imagery paradigm with auditory playback perturbation, we provide direct evidence suggesting that a gain control mechanism and dynamic time and frequency windows cooperatively govern the spectrotemporal comparison between internal prediction and external stimulation in speech. These findings demonstrate the cognitive mechanisms mediating self-monitoring and feedback control.

Acknowledgments

We thank Jeff Walker for his excellent technical support. This study was supported by MURI ARO 54228-LS-MUR and NIH 2R01DC 05660.

REFERENCES

Antipova EA, Purdy SC, Blakeley M, Williams S. Effects of altered auditory feedback (AAF) on stuttering frequency during monologue speech production. Journal of Fluency Disorders. 2008; 33:274–290. [PubMed: 19328980]

Behroozmand R, Karvelis L, Liu H, Larson CR. Vocalization-induced enhancement of the auditory cortex responsiveness during voice F0 feedback perturbation. Clinical Neurophysiology. 2009; 120:1303–1312. [PubMed: 19520602]

Behroozmand R, Liu H, Larson CR. Time-dependent neural processing of auditory feedback during voice pitch error detection. Journal of Cognitive Neuroscience. 2011; 23:1205–1217. [PubMed: 20146608]

Berger CC, Ehrsson HH. Mental imagery changes multisensory perception. Current Biology. 2013; 23:1367–1372. [PubMed: 23810539]

Black JW. The effect of delayed side-tone upon vocal rate and intensity. Journal of Speech & Hearing Disorders. 1951; 16:56–61.

Blakemore SJ, Wolpert DM, Frith CD. Central cancellation of self-produced tickle sensation. Nature Neuroscience. 1998; 1:635–640.

Blakemore SJ, Wolpert DM, Frith CD. Why can't you tickle yourself? NeuroReport. 2000; 11:R11. [PubMed: 10943682]

Boersma, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.. Paper presented at the Proceedings of the Institute of Phonetic Sciences; 1993.

Burke BD. Variables affecting stutterer's initial reactions to delayed auditory feedback. Journal of Communication Disorders. 1975; 8:141–155. [PubMed: 803158]

Chang EF, Niziolek CA, Knight RT, Nagarajan SS, Houde JF. Human cortical sensorimotor network underlying feedback control of vocal pitch. Proceedings of the National Academy of Sciences, U.S.A. 2013; 110:2653–2658.

Christoffels IK, van de Ven V, Waldorp LJ, Formisano E, Schiller NO. The sensory consequences of speaking: Parametric neural cancellation during speech in auditory cortex. PLoS One. 2011; 6:e18307. [PubMed: 21625532]

Davelaar EJ, Tian X, Weidemann CT, Huber DE. A habituation account of change detection in same/different judgments. Cognitive, Affective & Behavioral Neuroscience. 2011; 11:608–626.

de Cheveigné A, Simon JZ. Denoising based on time-shift PCA. Journal of Neuroscience Methods. 2007; 165:297–305. [PubMed: 17624443]

Dechent P, Merboldt KD, Frahm J. Is the human primary motor cortex involved in motor imagery? Brain Research, Cognitive Brain Research. 2004; 19:138–144. [PubMed: 15019710]

Deiber MP, Ibañez V, Honda M, Sadato N, Raman R, Hallett M. Cerebral processes related to visuomotor imagery and generation of simple finger movements studied with positron emission tomography. Neuroimage. 1998; 7:73–85. [PubMed: 9571132]

Desmurget M, Sirigu A. A parietal-premotor network for movement intention and motor awareness. Trends in Cognitive Sciences. 2009; 13:411–419. [PubMed: 19748304]

Ehrsson HH, Geyer S, Naito E. Imagery of voluntary movement of fingers, toes, and tongue activates corresponding body-part-specific motor representations. Journal of Neurophysiology. 2003; 90:3304–3316. [PubMed: 14615433]

Eliades SJ, Wang X. Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. Journal of Neurophysiology. 2003; 89:2194. [PubMed: 12612021]

Eliades SJ, Wang X. Dynamics of auditory-vocal interaction in monkey auditory cortex. Cerebral Cortex. 2005; 15:1510. [PubMed: 15689521]

Eliades SJ, Wang X. Neural substrates of vocalization feedback monitoring in primate auditory cortex. Nature. 2008; 453:1102–1106. [PubMed: 18454135]

Fairbanks G. Selective vocal effects of delayed auditory feedback. Journal of Speech & Hearing Disorders. 1955; 20:333–346. [PubMed: 13272227]

Gerardin E, Sirigu A, Lehericy S, Poline J-B, Gaymard B, Marsault C, et al. Partially overlapping neural networks for real and imagined hand movements. Cerebral Cortex. 2000; 10:1093–1104. [PubMed: 11053230]

Greenlee JD, Behroozmand R, Larson CR, Jackson AW, Chen F, Hansen DR, et al. Sensory-motor interactions for vocal pitch monitoring in non-primary human auditory cortex. PloS One. 2013; 8:e60783. [PubMed: 23577157]

Grush R. The emulation theory of representation: Motor control, imagery, and perception. Behavioral and Brain Sciences. 2004; 27:377–396. [PubMed: 15736871]

Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. Brain and Language. 2006; 96:280–301. [PubMed: 16040108]

Hanakawa T, Immisch I, Toma K, Dimyan MA, Van Gelderen P, Hallett M. Functional properties of brain areas associated with motor execution and imagery. Journal of Neurophysiology. 2003; 89:989–1002. [PubMed: 12574475]

Hargrave S, Kalinowski J, Stuart A, Armson J, Jones K. Effect of frequency-altered feedback on stuttering frequency at normal and fast speech rates. Journal of Speech, Language and Hearing Research. 1994; 37:1313.

Hickok G. Computational neuroanatomy of speech production. Nature Reviews Neuroscience. 2012; 13:135–145.

Hickok G, Houde J, Rong F. Sensorimotor integration in speech processing: Computational basis and neural organization. Neuron. 2011; 69:407–422. [PubMed: 21315253]

Hickok G, Poeppel D. The cortical organization of speech processing. Nature Reviews Neuroscience. 2007; 8:393–402.

Houde JF, Nagarajan SS. Speech production as state feedback control. Frontiers in Human Neuroscience. 2011; 5 Article 82.

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: An MEG study. Journal of Cognitive Neuroscience. 2002; 14:1125–1138. [PubMed: 12495520]

Howell P, Archer A. Susceptibility to the effects of delayed auditory feedback. Perception & Psychophysics. 1984; 36:296–302. [PubMed: 6522222]

Howell, P.; El-Yaniv, N.; Powell, DJ. Factors affecting fluency in stutterers when speaking under altered auditory feedback.. In: Peters, HFM.; Hulstijin, W., editors. Speech motor dynamics in stuttering. Springer-Verlag; New York: 1987. p. 361-369.

Huber DE, Tian X, Curran T, O'Reilly RC, Woroch B. The dynamics of integration and separation: ERP, MEG, and neural network studies of immediate repetition effects. Journal of Experimental Psychology: Human Perception and Performance. 2008; 34:1389–1416. [PubMed: 19045982]

Jeannerod M. The representing brain: Neural correlates of motor intention and imagery. Behavioral and Brain Sciences. 1994; 17:187–202.

Jeannerod M. Mental imagery in the motor context. Neuropsychologia. 1995; 33:1419–1432. [PubMed: 8584178]

Kalinowski J, Armson J, Stuart A, Gracco VL. Effects of alterations in auditory feedback and speech rate on stuttering frequency. Language and Speech. 1993; 36:1–16. [PubMed: 8345771]

Kalinowski J, Stuart A, Sark S, Armson J. Stuttering amelioration at various auditory feedback delays and speech rates. International Journal of Language & Communication Disorders. 1996; 31:259–269.

Kalman RE. A new approach to linear filtering and prediction problems. Journal of Basic Engineering. 1960; 82:35–45.

Katseff S, Houde J, Johnson K. Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? Language and Speech. 2012; 55:295–308. [PubMed: 22783636]

Liu H, Meshman M, Behroozmand R, Larson CR. Differential effects of perturbation direction and magnitude on the neural processing of voice pitch feedback. Clinical Neurophysiology. 2011; 122:951–957. [PubMed: 20869305]

Lotze M, Montoya P, Erb M, Hulsmann E, Flor H, Klose U, et al. Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fMRI study. Journal of Cognitive Neuroscience. 1999; 11:491–501. [PubMed: 10511638]

Luo H, Tian X, Song K, Zhou K, Poeppel D. Neural response phase tracks how listeners learn new acoustic representations. Current Biology. 2013; 23:968–974. [PubMed: 23664974]

Max L, Guenther FH, Gracco VL, Ghosh SS, Wallace ME. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. Contemporary Issues in Communication Science and Disorders. 2004; 31:105–122.

Meister IG, Krings T, Foltys H, Boroojerdi B, Müller M, Töpper R, et al. Playing piano in the mind: An fMRI study on music imagery and performance in pianists. Brain Research, Cognitive Brain Research. 2004; 19:219–228. [PubMed: 15062860]

Miall RC, Wolpert DM. Forward models for physiological motor control. Neural Networks. 1996; 9:1265–1279. [PubMed: 12662535]

Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication. 1990; 9:453–467.

Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. Attention, Perception, & Psychophysics. 1996; 58:351–362.

Naylor RV. A comparative study of methods of estimating the severity of stuttering. Journal of Speech and Hearing Disorders. 1953; 18:30. [PubMed: 13053561]

Numminen J, Salmelin R, Hari R. Subject's own speech reduces reactivity of the human auditory cortex. Neuroscience Letters. 1999; 265:119–122. [PubMed: 10327183]

Poeppel D. The analysis of speech in different temporal integration windows: Cerebral lateralization as “asymmetric sampling in time.”. Speech Communication. 2003; 41:245–255.

Roberts TPL, Ferrari P, Stufflebeam SM, Poeppel D. Latency of the auditory evoked neuromagnetic field components: Stimulus dependence and insights toward perception. Journal of Clinical Neurophysiology. 2000; 17:114–129. [PubMed: 10831104]

Scheerer N, Behich J, Liu H, Jones J. ERP correlates of the magnitude of pitch errors detected in the human voice. Neuroscience. 2013; 240:176–185. [PubMed: 23466810]

Scott M. Corollary discharge provides the sensory content of inner speech. Psychological Science. 2013; 24:1824–1830. [PubMed: 23846719]

Shearer W. Self-masking effects from live and recorded vowels. Journal of Auditory Research. 1978; 18:213. [PubMed: 755818]

Shuster LI, Durrant JD. Toward a better understanding of the perception of self-produced speech. Journal of Communication Disorders. 2003; 36:1–11. [PubMed: 12493635]

Sommer MA, Wurtz RH. A pathway in primate brain for internal monitoring of movements. Science. 2002; 296:1480. [PubMed: 12029137]

Sommer MA, Wurtz RH. Influence of the thalamus on spatial visual processing in frontal cortex. Nature. 2006; 444:374–377. [PubMed: 17093408]

Sommer MA, Wurtz RH. Brain circuits for the internal monitoring of movements. Annual Review of Neuroscience. 2008; 31:317–338.

Stuart A, Kalinowski J, Armson J, Stenstrom R, Jones K. Fluency effect of frequency alterations of plus/minus one-half and one-quarter octave shifts in auditory feedback of people who stutter. Journal of Speech, Language and Hearing Research. 1996; 39:396.

Tian X, Huber DE. Measures of spatial similarity and response magnitude in MEG and scalp EEG. Brain Topography. 2008; 20:131–141. [PubMed: 18080180]

Tian X, Huber DE. Playing “duck duck goose” with neurons change detection through connectivity reduction. Psychological Science. 2013; 24:819–827. [PubMed: 23572279]

Tian X, Poeppel D. Mental imagery of speech and movement implicates the dynamics of internal forward models. Frontiers in Psychology. 2010; 1 Article 166.

Tian X, Poeppel D. Mental imagery of speech: Linking motor and sensory systems through internal simulation and estimation. Frontiers in Human Neuroscience. 2012; 6 Article 314.

Tian X, Poeppel D. The effect of imagination on stimulation: The functional specificity of efference copies in speech processing. Journal of Cognitive Neuroscience. 2013; 25:1020–1036. [PubMed: 23469885]

Tian X, Poeppel D, Huber DE. TopoToolbox: Using sensor topography to calculate psychologically meaningful measures from event-related EEG/MEG. Computational Intelligence and Neuroscience. 2011 doi:10.1155/2011/674605.

Tourville JA, Reilly KJ, Guenther FH. Neural mechanisms underlying auditory feedback control of speech. Neuroimage. 2008; 39:1429–1443. [PubMed: 18035557]

van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007; 45:598–607. [PubMed: 16530232]

Vaxes AJ. Delayed auditory feedback. Psychological Bulletin. 1963; 60:213–232. [PubMed: 14002534]

Ventura M, Nagarajan S, Houde J. Speech target modulates speaking induced suppression in auditory cortex. BMC Neuroscience. 2009; 10:58. [PubMed: 19523234]

Webster RL, Schumacher SJ, Lubker BB. Changes in stuttering frequency as a function of various intervals of delayed auditory feedback. Journal of Abnormal Psychology. 1970; 75:45. [PubMed: 5416040]

Wolpert DM, Ghahramani Z. Computational principles of movement neuroscience. Nature Neuroscience. 2000; 3:1212–1217.

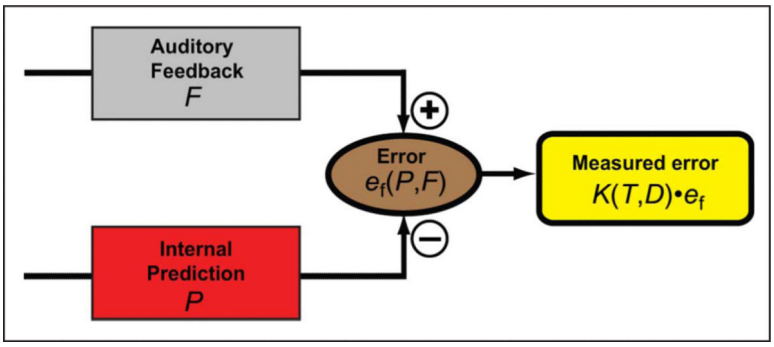

Figure 1.

Proposed model: Kalman gain modulates errors terms (e_f) between internal prediction and external feedback. The Kalman gain (K; Equation 1), which is a function of spectral (D) and temporal (T) differences between prediction and feedback, actively modulates the magnitude of error terms that are formed by comparing the prediction (P) and feedback (F).

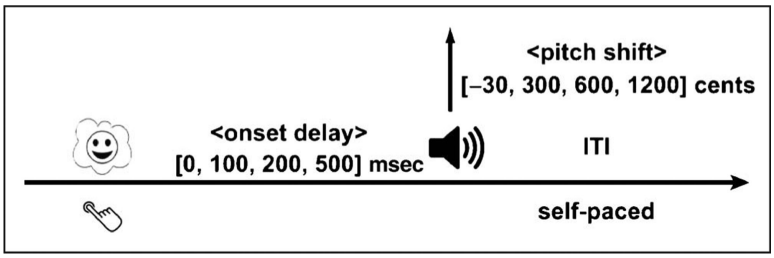

Figure 2.

Schematic description of the experimental design. Participants press a button at the beginning of articulation imagery of the vowel /a/. The prerecorded individual vocalization of /a/ was manipulated and formed four levels of pitch-shifted playback that was presented at four levels of delays after the button press.

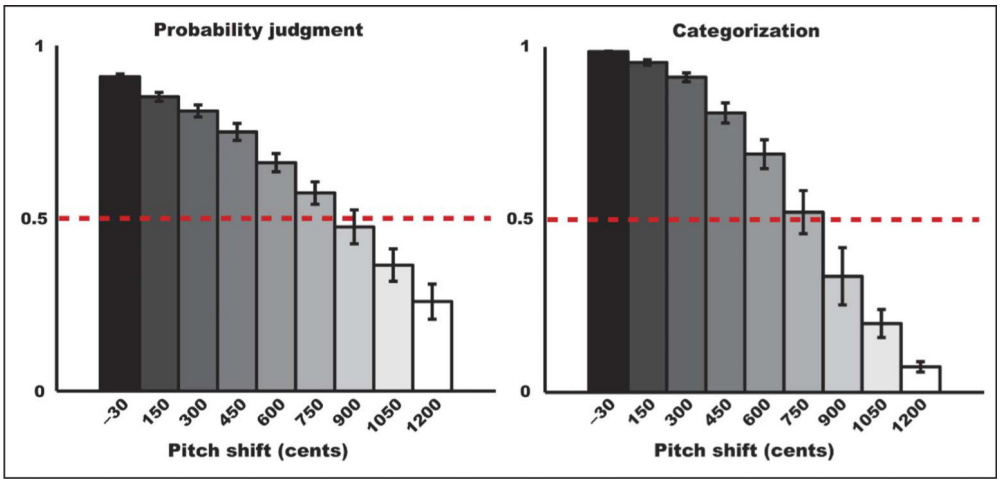

Figure 3.

Behavioral results of pitch range assessment. The subjective probability judgment of pitch range in one's normal conversation is plotted as a function of pitch shift. Left: Results in Behavioral Experiment 1 using probability judgment. Right: Results in Behavioral Experiment 2 using categorization task. The red dash lines represented the chance level.

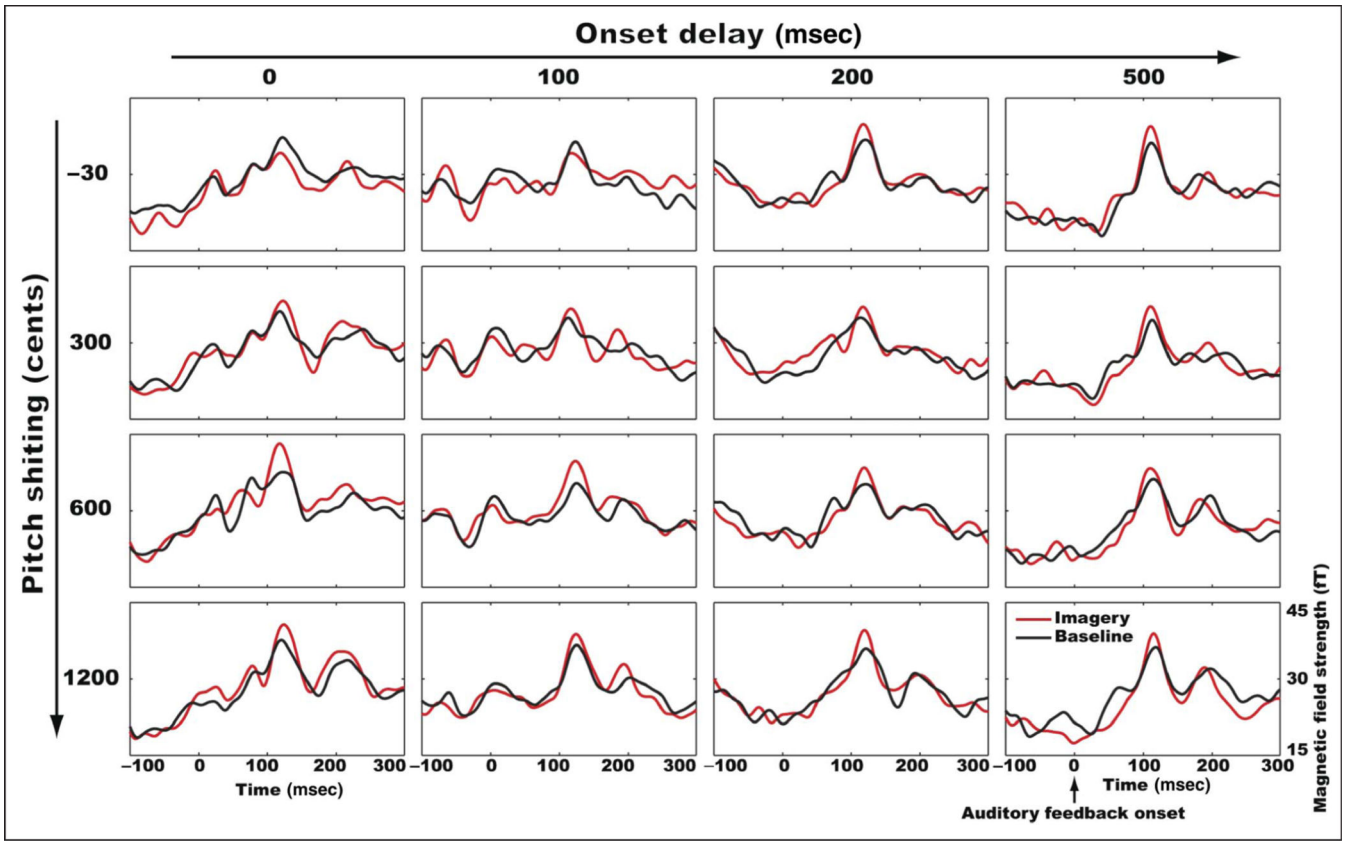

Figure 4.

Waveform responses to playback in all conditions during experimental and baseline runs. RMS waveform responses to auditory feedback in all conditions were plotted. Two lines were included in each subplot, with the red line representing the response during experimental run and the black line for baseline run. Subplots were arranged vertically as the level of pitch shift increases and horizontally as the level of onset delay increases.

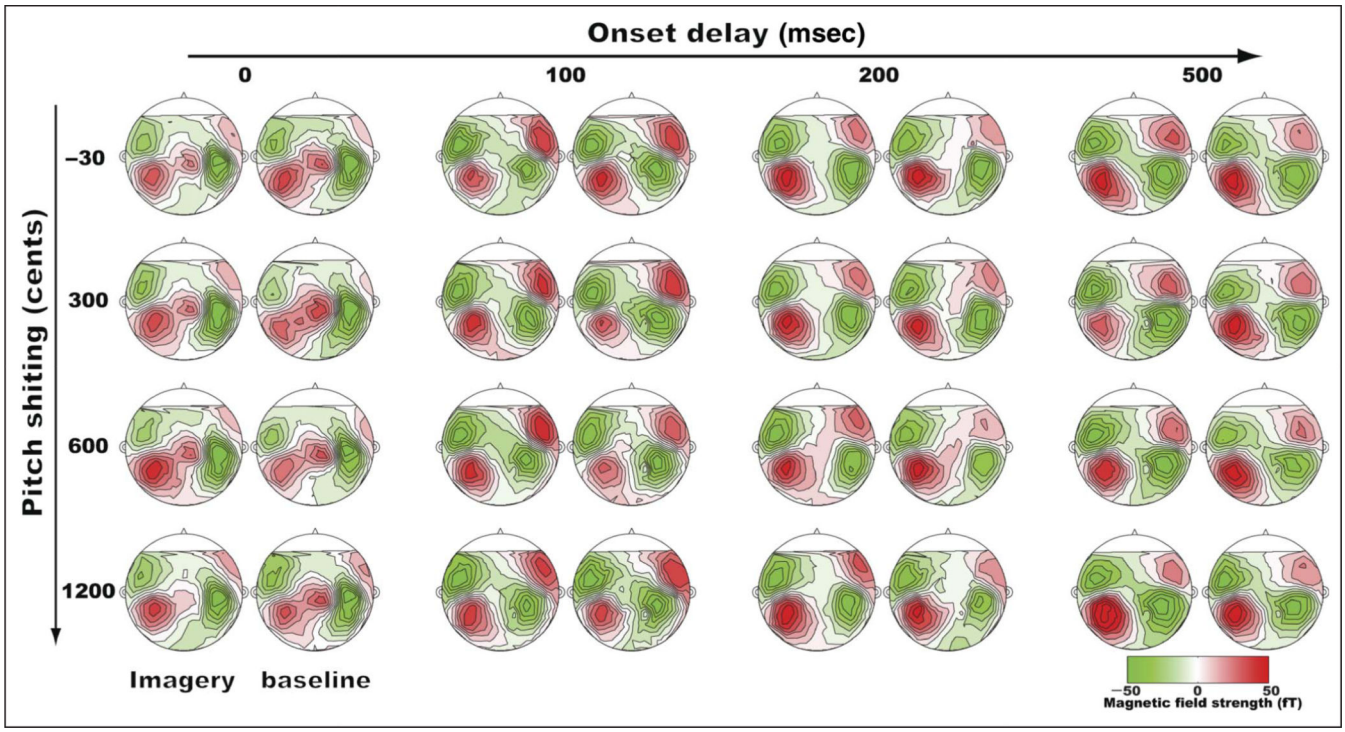

Figure 5.

Topographies of auditory M100 responses to playback in all conditions during experimental and baseline runs. Topographies are grouped in a pair for each condition, with the one in experimental run on left and baseline run on right. The layout of topographies for all conditions is identical as Figure 4.

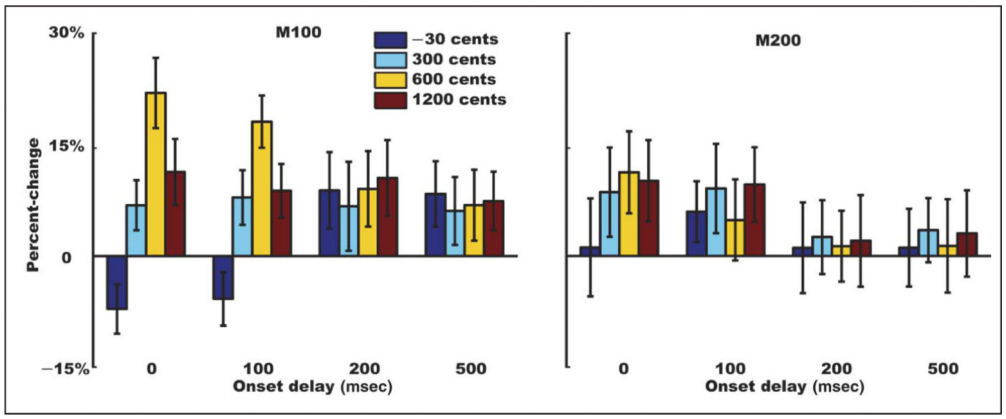

Figure 6.

Response differences in M100 (left) and M200 (right) components as functions of pitch shift and onset delay. The relative magnitude changes, calculated by subtracting the auditory response in baseline run from the one in experimental run, are plotted as a function of pitch shift and grouped by each level of onset delay.

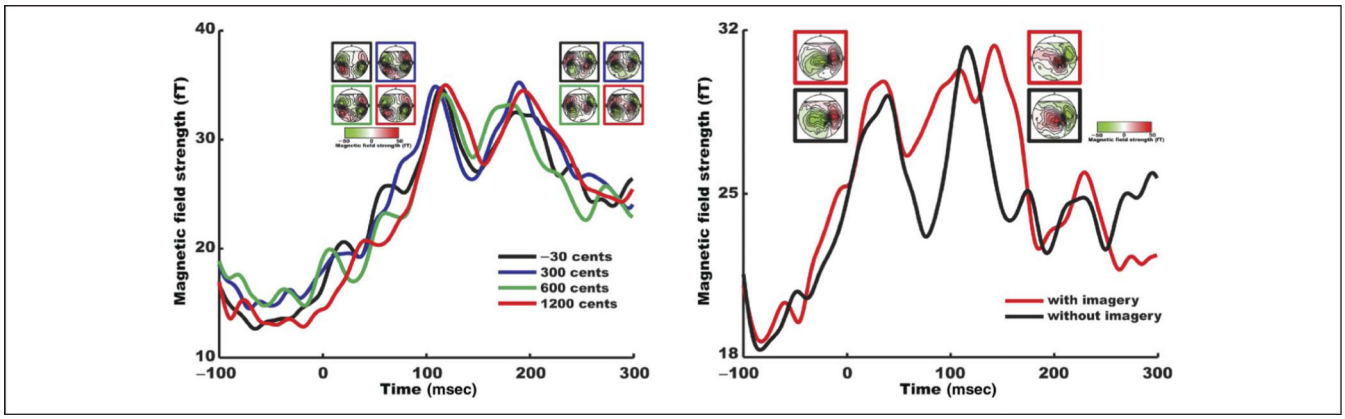

Figure 7.

Waveform responses and topographies of control runs. Left: Waveform responses and topographies of four different pitch shifted sounds in auditory control. RMS waveforms to four pitch shifted sounds are plotted in different colors. M100 and M200 components were inserted beside the waveforms at corresponding latencies. Color boxes that indicate topography of responses to each sound use the same color code as waveforms. Right: Waveforms responses and topographies of button press in motor control runs. RMS waveforms to button press responses are plotted for runs with (red) and without (black) articulation imagery. Two response components, presumably mediating button press and release, are plotted near the corresponding latency, with the surrounding boxes being identical color code as waveforms.