nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-018-0305-8

Imagined speech influences perceived loudness of sound

Xing Tian[1,2,3*], Nai Ding[4,5,6], Xiangbin Teng[7,8], Fan Bai[1,2,3] and David Poeppel[7,8]

The way top-down and bottom-up processes interact to shape our perception and behaviour is a fundamental question and remains highly controversial. How early in a processing stream do such interactions occur, and what factors govern such interactions? The degree of abstractness of a perceptual attribute (for example, orientation versus shape in vision, or loudness versus sound identity in hearing) may determine the locus of neural processing and interaction between bottom-up and internal information. Using an imagery-perception repetition paradigm, we find that imagined speech affects subsequent auditory perception, even for a low-level attribute such as loudness. This effect is observed in early auditory responses in magnetoencephalography and electroencephalography that correlate with behavioural loudness ratings. The results suggest that the internal reconstruction of neural representations without external stimulation is flexibly regulated by task demands, and that such top-down processes can interact with bottom-up information at an early perceptual stage to modulate perception.

There is considerable evidence that top-down processes can change our perception of the same physical stimulus[1,2]. For example, you are more likely to hear your phone ring when you are expecting a call, and you are less likely to hear a sound when you are busy reading[3]. Such top-down expectancy modulation effects on perception might occur in a specific way: a similar internal representation to the representation elicited by overt perception can be induced via top-down processes without physical stimulation, and by hypothesis this internally generated representation interacts with bottom-up incoming sensory information and thereby modulates perception. (Hereafter in this paper, neural and psychological processes that are triggered by external physical stimulation are referred to as 'bottom-up' processes. In contrast, neural and psychological processes that are internally induced without any external stimulation are referred to as 'top-down' processes.) Indeed, mental imagery and memory have been proposed to include the process of mental representation 'reconstruction' without physical stimuli (for example, refs [4–7]) and influence perception. However, the evidence diverges in unanticipated ways when considering the level of neural processing between the visual and auditory domains[8,9]. For example, numerous neuroimaging studies suggest that the primary visual cortex is principally involved in visual imagery[10–12]. Although significantly fewer auditory imagery neuroimaging studies have been carried out, most of them do not observe the primary auditory cortex in auditory imagery (for example, refs [13,14]). The inconsistent functional and anatomical results between perceptual domains have held back theoretical advances on how the brain integrates top-down information with information from external stimulation to shape perception.

The degree of abstractness of a perceptual attribute (for example, orientation versus shape) may determine the level of neural processing and the interaction between bottom-up information and internal information[15]. On this hypothesis, the inconsistent observations between visual and auditory domains could be caused by the focus on different levels of attributes. For example, in vision, recent behavioural experiments demonstrate that imagery can affect all levels of perception, from high-level spatial configurations[16] to lower-level attributes such as orientation[17,18]—and even muscle control and pupil contraction[19]. In contrast, behavioural studies in the auditory domain typically focus on higher-level attributes, such as syllable-level representation[20].

Here, we manipulate the level of abstraction of a perceptual attribute in the auditory domain (1) to examine electrophysiologically whether we observe functionally early perceptual activation without physical stimulation, as is observed in the visual domain, and (2) to investigate whether perception of a basic auditory attribute can be modulated by the reconstruction of early perceptual responses. Specifically, we investigate whether mental imagery affects auditory perception, even for a basic auditory attribute such as loudness. We use behavioural tasks, magnetoencephalography (MEG) and electroencephalography (EEG) to test whether the level of perceptual analysis determines the level in the processing hierarchy that mediates the neural and perceptual modulation. That is, we test whether 'thought' influences loudness perception and we ask whether early auditory neural responses from putative primary auditory cortex underlie the modulatory effect of mental imagery on loudness perception.

Stimulus repetition is a powerful paradigm for investigating mental representation: mental and neural processes mediating a specific stimulus class can be identified by response decreases when the same stimulus repeats[21]. Here, we pair a top-down initiated process (imagined speech, at different loudness levels) with a bottom-up process (loudness perception of overt speech) in a repetition design to characterize the interaction between imagery and the subsequent loudness judgement. We hypothesized that internally constructing a loudness percept during imagery activates the

[1]New York University Shanghai, Shanghai, China. [2]Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China. [3]New York University-East China Normal University Institute of Brain and Cognitive Science at New York University Shanghai, Shanghai, China. [4]College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Zhejiang, China. [5]Key Laboratory for Biomedical Engineering, Ministry of Education, Zhejiang University, Zhejiang, China. [6]State Key Laboratory of Industrial Control Technology, Zhejiang University, Zhejiang, China. [7]Department of Psychology, New York University, New York, NY, USA. [8]Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany. *e-mail: xing.tian@nyu.edu

putative representation of loudness in early auditory regions that mediate actual loudness perception. If the same neural representation activates again during subsequent perception of an overt sound after the preceding imagery, and if imagining speaking loudly activates the representation to a greater degree than imagining speaking softly, imagining speaking loudly will tax more neural resources that mediate the common representation in imagery and perception, and hence reduce the subsequent perceptual response more. Therefore, we predicted that the loudness ratings relating to the overt probe sounds would be lower after imagining speaking loudly than after imagining speaking softly. Moreover, early auditory cortical responses should have lower amplitudes in the loud versus soft conditions, and the hypothesized neural sources of the top-down modulation effects should localize to early auditory cortical areas.

Results

Behavioural experiment 1 (BE1): loudness judgement after imagined speech. In BE1, participants were asked to imagine speaking the syllable 'da' either loudly or softly four times before making a loudness judgement for an overt auditory stimulus of the same syllable (Fig. 1a). The results in BE1 showed that loudness ratings were consistently lower when participants imagined speaking loudly than when they imagined speaking softly (Fig. 1b). A repeated-measures two-way analysis of variance (ANOVA) was carried out with the factors of imagined loudness and sound intensity. The main effect of sound intensity was significant ($F_{4,60}=143.72$, $P<0.001$). More importantly, the main effect of imagined loudness was also significant ($F_{1,60}=5.57$, $P=0.032$). However, the interaction was not significant ($F_{4,60}=2.17$, $P=0.084$). This suggests that the higher the 'internal volume' during imagined speech, the softer the perceived loudness for subsequent sounds.

Behavioural experiment 2 (BE2): loudness judgement after imagined speech in noise. In BE2, white noise was introduced during the imagery period. If the observed modulation effect is due to imagery itself, the background noise during imagery would disrupt the observed modulation effects of imagery. A repeated-measures two-way ANOVA with the factors of imagined loudness and sound intensity showed that the main effect of sound intensity was significant ($F_{4,60}=506.7$, $P<0.001$). However, neither the main effect of imagery type ($F<1$) nor the interaction ($F_{4,60}=1.75$, $P=0.15$) was significant (Fig. 1c). That is, the modulation effect of imagery on subsequent loudness perception was disrupted by the background noise during imagery. These results support the hypothesis that the observed modulation effect is a low-level effect during imagery.

Behavioural experiment 3 (BE3): loudness judgement after different repetitions of imagined speech. In BE3, we manipulated the number of times participants were required to imagine speaking the syllable 'da' (Fig. 2a). If the modulation effect is a lower-level effect of imagery, it would reflect the accumulation over time. That is, the more times imagery is executed before hearing the overt sound, the greater the modulation effects. As shown in Fig. 2b, the main effect of imagery times is significant ($F_{2,30}=3.53$, $P=0042$). Post-hoc dependent *t*-tests revealed that rating differences between loud and soft conditions became more negative (rating less in loud) when imagery was executed 5 or 6 times, compared with imagery executed 3 or 4 times ($t_{15}=-2.23$, $P=0.041$). The rating difference decrease was marginally significant when comparing between imagery executed 3 or 4 times and imagery executed 1 or 2 times ($t_{15}=-2.08$, $P=0.055$). No differences were found in the rating difference between imagery executed 5 or 6 times and imagery executed 3 or 4 times ($t<1$). These results suggest that the modulation effect

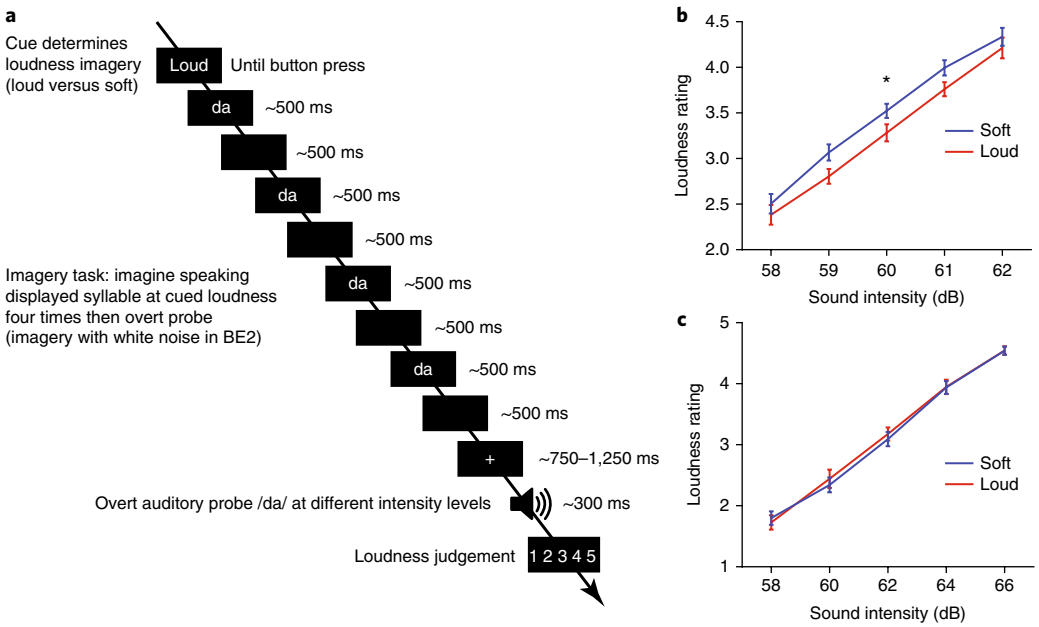

Fig. 1 | Experimental procedure and behavioural results for BE1 and BE2. a, Example trial for the loud condition. Participants imagined saying the syllable 'da' four times, paced by the visual cue 'da' (presented every second). Participants imagined saying the syllable loudly or softly according to the visual cue presented at the beginning of a trial. When the participants imagined speaking 'da', no sound was presented in BE1, whereas white noise was presented in BE2. A fixation cross appeared after they finished four iterations of imagined speech and was followed by the presentation of the same syllable (pre-recorded individually). One of five intensity levels of the auditory syllable was randomly chosen on a trial, and participants were asked to provide a loudness judgement for this auditory stimulus. **b**, Loudness rating results for BE1. The loudness ratings subsequent to imagined soft speech (blue) were higher than those following imagined loud speech (red), and this was consistent across different levels of intensity for the auditory stimuli. **c**, Loudness rating results for BE2. The physically present white noise abolished the imagery effect on the loudness rating. Loudness ratings in the loud (red) and soft (blue) conditions were not different at any level of intensity. $^*P<0.05$. Error bars represent s.e.m. ($n=16$).

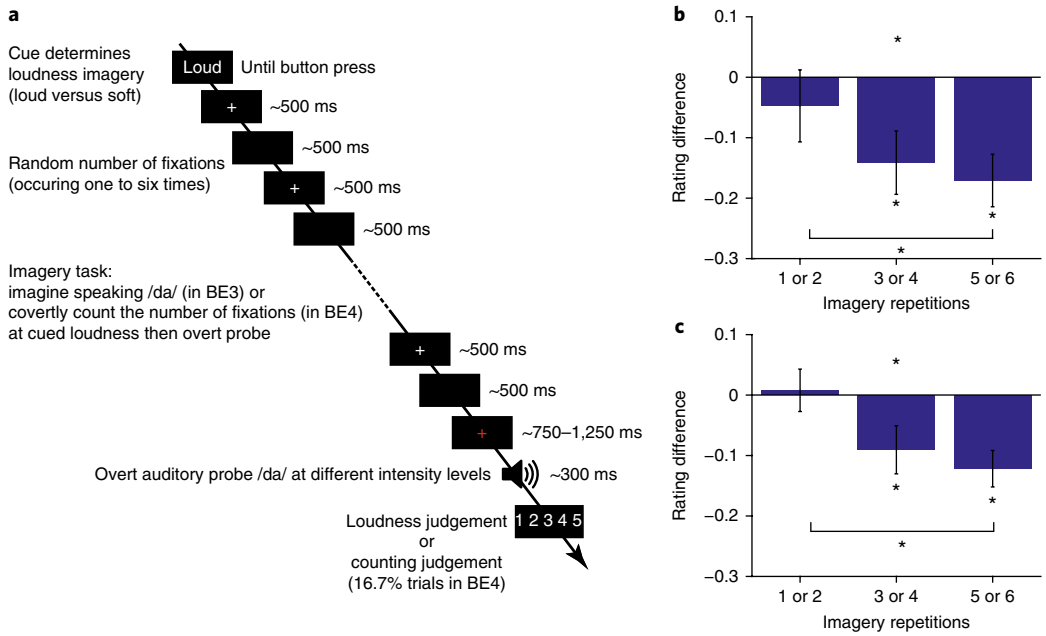

Fig. 2 | Experimental procedure and behavioural results for BE3 and BE4. a, Example trial for the loud condition for BE3 and BE4. Participants imagined saying the syllable 'da', paced by the visual fixation (presented every second) in BE3, or imagined counting the number of fixations in BE4. The occurrence of fixations in a trial was random and between one and six times. Participants performed the imagery task loudly or softly according to the visual cue presented at the beginning of each trial. A red fixation cross appeared to indicate the end of imagery and was followed by the presentation of the auditory syllable 'da' (pre-recorded individually). One of five intensity levels of the auditory syllable was randomly chosen for each trial, and participants were asked to provide a loudness judgement for this auditory stimulus. For BE4, in 16.7% of trials, participants were required to make judgement on how many times they counted the fixation (counting judgement). **b**, Loudness rating results for BE3. The loudness rating differences between loud and soft were plotted as a function of imagery repetitions. The rating differences (loud minus soft) became more negative as the imagery times increased, as indicated by the significant differences in ANOVA test results among different numbers of imagery repetitions, paired *t*-test results between 1 or 2 and 5 or 6 imagery repetitions, and one-sample test results for 3 or 4 and 5 or 6 imagery repetitions against 0. **c**, Loudness rating results for BE4. Same analysis tests as in **b**. *$P<0.05$. Error bars represent s.e.m. ($n=16$).

of imagery on loudness perception is a function of imagery time. That is, the more times imagery is executed, the greater the strength of modulation. One-sample *t*-tests revealed that the rating difference between loud and soft was significantly smaller than 0 when imagery was executed 3 or 4 times ($t_{15}=-2.70$, $P=0.017$), as well as when imagery was executed 5 or 6 times ($t_{15}=-3.93$, $P=0.001$), but for imagery executed 1 or 2 times the rating difference was not different from 0 ($t<1$). These results show that the modulation effects need around three to four executions of imagery, which is consistent with our results in the main behavioural experiment. These control results support that the observed modulation effect is a low-level effect during imagery.

Behavioural experiment 4 (BE4): loudness judgement after imagined counting. In BE4, an imagined counting task was used. Moreover, the counting judgement task was intermixed with a loudness judgement task. For the counting judgement task, the average accuracy was 0.92, and the s.e.m. was 0.02. The significantly-above-chance (accuracy of 0.5) performance in the counting task suggests that participants were actively engaged in the imagery task. For the loudness judgement task (Fig. 2c), the main effect of imagery times was significant ($F_{2,30}=3.56$, $P=0.041$). Post-hoc dependent *t*-tests revealed that rating differences between the loud and soft conditions became more negative (rating less in loud) when imagery was executed 5 or 6 times, compared with when imagery was executed 3 or 4 times ($t_{15}=-2.77$, $P=0.014$). The rating difference decrease was marginally significant when comparing between imagery executed 3 or 4 times and imagery executed 1 or 2 times ($t_{15}=-1.81$, $P=0.09$). No differences were found in the rating difference between imagery executed 5 or 6 times and imagery executed 3 or 4 times ($t<1$). One-sample *t*-tests revealed that the rating difference between loud and soft was significantly smaller than 0 when imagery was executed 3 or 4 times ($t_{15}=-2.70$, $P=0.017$), as well as when imagery was executed 5 or 6 times ($t_{15}=-3.93$, $P=0.001$), but when imagery was executed 1 or 2 times, the rating difference was not different from 0 ($t<1$). These results are consistent with those obtained in BE3, and further suggest that the modulation effects of imagery on loudness perception are not content specific.

MEG experiment: imagined speech induces loudness neural adaptation. The MEG results (Fig. 3) were consistent with the behavioural data. The neural response topographies were similar between the soft, loud and no-imagery conditions for the M100 and M200 neural responses that occurred around 100 ms and 200 ms, respectively, after the auditory stimuli were presented (Fig. 3a). The response pattern similarity between conditions was quantified by the angle test[22,23], which suggested that the topographies of auditory responses among the 3 conditions were quantitatively similar (for no-imagery versus soft, M100: $t_{15}=5.23$, $P<0.001$; M200: $t_{15}=3.18$, $P=0.006$; for no-imagery versus loud, M100: $t_{15}=6.85$, $P<0.001$, M200: $t_{15}=2.60$, $P=0.02$; for soft versus loud, M100: $t_{15}=5.37$, $P<0.001$, M200: $t_{15}=2.13$, $P=0.05$). Therefore, the responses probably arose from similar neural sources, and the following results of response magnitude testing were free from the potential confounds of source distribution differences.

Critically, the response power to the auditory probes was modulated by the preceding imagery task (Fig. 3a). Specifically, the magnitude of the M100 component—an early cortical response

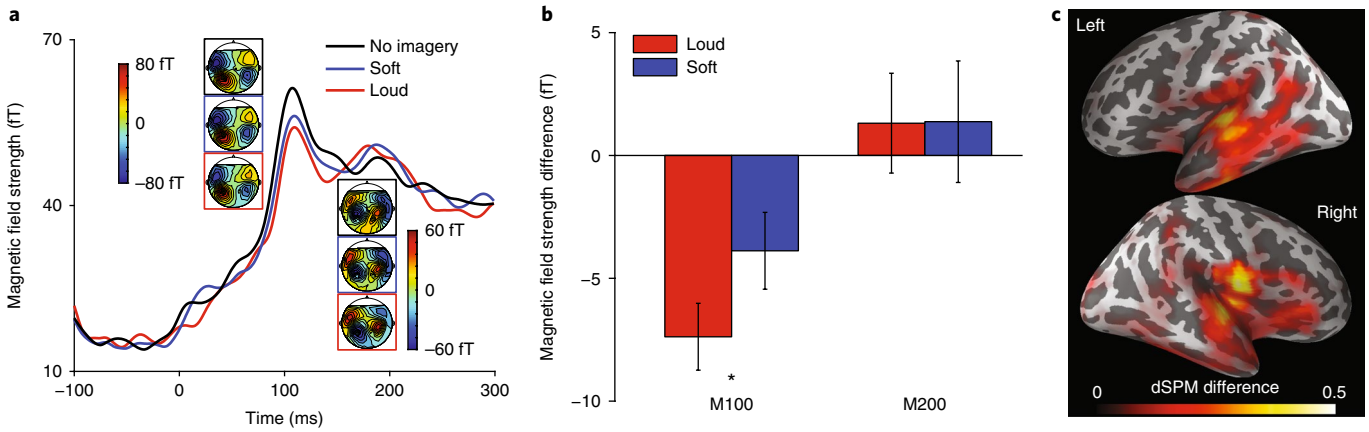

Fig. 3 | Neural adaptation results in the MEG experiment. a, Response and topography results. Each line represents the global response power across all MEG sensors, with the red line depicting the loud condition, the blue line the soft condition and the black line the no-imagery condition. Two clear peaks—one around 100 ms (M100) and another around 200 ms (M200)—after the syllable onset were observed. The response topographies at each peak time are shown in coloured boxes near each peak, using the same colour coding to represent each condition. **b**, Response changes (relative to responses in the no-imagery condition) in loud (red) and soft (blue) conditions. The MEG response magnitude was obtained by temporally averaging in a 25 ms time window centred at the individual early (M100) and late (M200) peak latencies observed in **a**. The M100 response for the loud condition was significantly smaller than for the soft condition, as indicated by a paired *t*-test. **c**, Group-averaged MEG source localization results. Bilateral early auditory cortices were activated more by the auditory syllable after soft imagined speech than that after loud imagined speech. Normalized response (dynamic statistical parametric mapping (dSPM)) differences between the soft and loud conditions (soft minus loud) are depicted. $^{*}P < 0.05$. Error bars represent s.e.m. ($n = 16$).

reflecting auditory perceptual analysis[24]—changed as a function of the imagery condition: the responses for no-imagery were largest, followed by the soft condition, while the smallest response amplitudes appeared in the loud condition. The modulatory effects of imagined speech loudness on the neural response are summarized in Fig. 3b by showing the response differences for soft versus loud relative to the responses for no imagery. Repeated-measures one-way ANOVAs were carried out for M100 and M200 separately. The main effect of imagined loudness was significant for the M100 responses ($F_{2,30} = 14.32$, $P < 0.001$). The subsequent pairwise *t*-tests revealed that the response magnitude in the soft condition was less than that in the no-imagery condition ($t_{15} = -2.49$, $P = 0.025$), and the response magnitude in the loud condition was less than that in the soft condition ($t_{15} = -2.93$, $P = 0.01$). However, the effect was not significant in the later auditory response (M200; $F < 1$). Source localization showed that the effects originated in bilateral auditory areas (Fig. 3c). Thus, loudness imagery modulated the early auditory cortical responses.

EEG experiment: imagined speech induces loudness neural adaptation, which correlates with behaviour. The results of the EEG experiment (Fig. 4) replicate the observations obtained in the behavioural and MEG experiments. The behavioural data in the EEG experiment showed that there was a main effect of imagined loudness (Fig. 4a; $F_{2,136} = 13.85$, $P < 0.001$). Further tests suggested that loudness ratings were consistently lower when participants imagined speaking softly than when no imagery was performed ($F_{1,68} = 6.87$, $P = 0.018$). Loudness ratings were also consistently lower when participants imagined speaking loudly than when they imagined speaking softly ($F_{1,68} = 11.39$, $P = 0.004$).

The EEG topographies were similar between the soft, loud and no-imagery conditions for the N1 and P2 components—EEG responses that occurred around 100 ms and 200 ms, respectively, after the presentation of auditory stimuli (these were at similar latencies to M100 and M200 observed in the MEG experiment) (Fig. 4b). The responses probably arose from similar neural sources. The response power, represented as the root-mean-square (RMS) of waveforms, changed as a function of the imagery condition (that is, the responses for the no-imagery condition were largest, followed by the soft condition, while the smallest response amplitudes appeared for the loud condition. The modulatory effects of imagined speech loudness on the neural response are summarized in Fig. 4c, which shows the normalized response changes for the soft and loud conditions relative to the no-imagery condition. The amplitude of early auditory responses (N1) for the loud condition was less than for the soft condition ($t_{17} = -2.78$, $P = 0.013$), the amplitude for the soft condition was less than for the no-imagery condition ($t_{17} = -3.15$, $P = 0.006$) and the amplitude for the loud condition was less than for the no-imagery condition ($t_{17} = -4.81$, $P < 0.001$). The effect was not significant in the later auditory response (P2; $t < 1$).

An analysis based on Bayes factors revealed that all Bayes factors of comparison between the evoked responses during the baseline period for the imagery (loud and soft) and no-imagery conditions favoured the null (for the comparison of loud and no imagery, scaled JZS Bayes factor = 3.43; for the comparison of soft and no imagery, scaled JZS Bayes factor = 2.06; for the comparison of loud and soft, scaled JZS Bayes factor = 3.98; for the comparison of no imagery and the average of loud and soft, scaled JZS Bayes factor = 2.75). All Bayes factors of induced responses in all frequency bands also favoured the null (for the comparison of loud and no imagery in the alpha band, scaled JZS Bayes factor = 1.39; in the beta band, scaled JZS Bayes factor = 1.07; in the low gamma band, scaled JZS Bayes factor = 4.00; for comparison of soft and no imagery in the alpha band, scaled JZS Bayes factor = 4.10; in the beta band, scaled JZS Bayes factor = 4.08; in the low gamma band, scaled JZS Bayes factor = 1.91; for comparison of loud and soft in the alpha band, scaled JZS Bayes factor = 1.27; in the beta band, scaled JZS Bayes factor = 1.28; in the low gamma band, scaled JZS Bayes factor = 1.36; for the comparison of no imagery and the average of loud and soft in the alpha band, scaled JZS Bayes factor = 2.70; in the beta band, scaled Bayes factor = 2.58; in the low gamma band, scaled JZS Bayes factor = 3.32). Although the short duration of baseline did not allow us to obtain direct evidence to rule out the possibility of activity differences in low-frequency

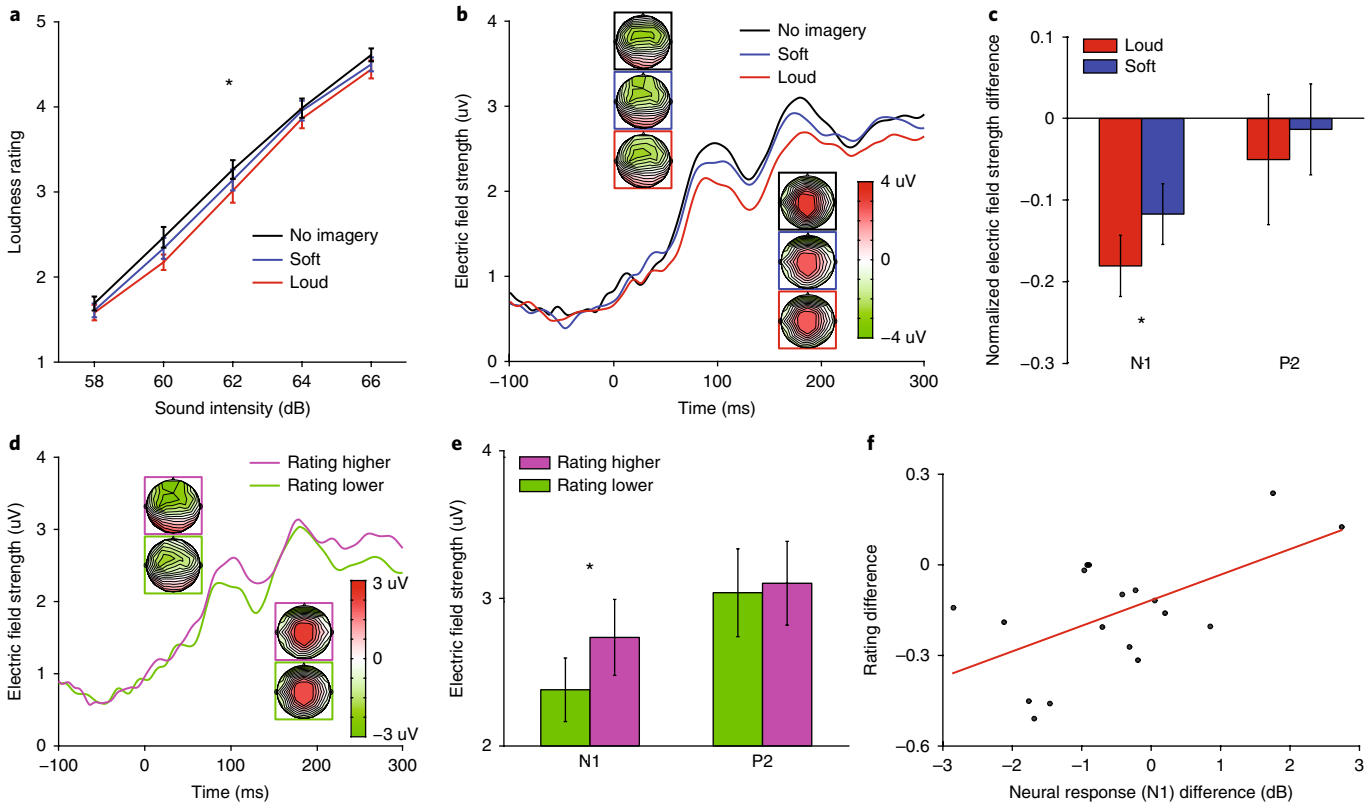

Fig. 4 | Results of the EEG experiment. a, Behavioural loudness ratings during the EEG experiment. The loudness ratings for the loud condition (red) were lower than those for the soft condition (blue), which were lower than those for no imagery (black). **b**, ERP time course and topography. RMS waveform (representing the square-root of the global response power) averaged across all EEG electrodes. Two peaks were observed in the ERP time course—one around 100 ms (N1) and the other around 200 ms (P2)—from auditory syllable onset. The response topographies at each peak time are shown in coloured boxes near each peak, using the same colour coding to represent each condition. **c**, Normalized response changes (relative to the no-imagery condition) in the loud (red) and soft (blue) conditions. The ERP response magnitude was obtained by temporally averaging in a 25 ms time window centred around the N1 and P2 peaks observed in **b**, and normalized by the responses for the no-imagery condition. The N1 response for the loud condition was significantly smaller than that for the soft condition, as indicated by a paired *t*-test. **d**, ERP time course and topography responses for different behavioural loudness ratings. These were obtained by averaging trials in the loud and soft conditions that had higher (purple) or lower (green) loudness ratings than the actual levels of sound intensity. The response topographies at each peak time are shown in coloured boxes near each peak, using the same colour coding as for the waveform responses. **e**, N1 and P2 response magnitudes for different behavioural loudness ratings. EEG response magnitudes were obtained by temporally averaging around the N1 and P2 peaks observed in **d**. The N1 response for rating higher trials was significantly larger than for rating lower trials, as indicated by a paired *t*-test. **f**, Results of correlation analysis. A significant positive correlation was observed between the behavioural rating decreases and the neural response decreases. $^*P < 0.05$. Error bars represent s.e.m. ($n = 18$).

bands during the baseline period, the Bayesian analysis results using the evoked responses suggest that these are less likely to be different across conditions. These results are consistent with the argument that there is no significant difference between the imagery and no-imagery conditions in the baseline period immediately before the stimulus and hence it cannot contaminate the modulation effects in the responses to the auditory stimuli.

Importantly, the modulation of imagery on neural responses to auditory stimuli correlated with behavioural ratings of loudness. As Fig. 4d,e demonstrates, in the loud and soft conditions, N1 responses were smaller when loudness was rated lower compared with the actual level of sound intensity versus when loudness was rated higher ($t_{17} = -2.34$, $P = 0.032$). The effect was not significant at P2. This suggests that a suppression of the N1 response associates directly with a decrease in loudness judgement. For the behavioural and neural response correlation analysis (Fig. 4f), there was a positive correlation between the loudness rating decreases and neural response N1 decreases (correlation coefficient $r_{16} = 0.58$, $P = 0.01$). Linear regression analysis revealed that the neural response N1 decreases predicted the behavioural loudness rating decreases (standardized coefficient $\beta = 0.085$, $R^2 = 0.34$, $t_{16} = 2.85$, $P = 0.01$). That is, the perceived loudness for subsequent sounds was correlated with the neural modulation of a preceding imagery process.

Discussion

Our results from four behavioural and two electrophysiological experiments cumulatively demonstrate that mental imagery can immediately and directly influence the auditory perceptual analysis of a basic, low-level sound attribute: loudness. The data show that top-down auditory imagery induces neural activity patterns similar to perceptual responses, and that such internally induced activation can modulate the perception of a low-level acoustic attribute. Our behavioural and electrophysiological results are consistent with the results of previous neuroimaging studies that demonstrate the recruitment of auditory cortices in auditory imagery[13,14,25], and that relevant auditory features are encoded and processed during imagery in specific auditory areas such as frequency[26] and categorical information[27]. Our results are also consistent with the results from EEG studies that show the reconstruction of representation during auditory imagery for higher-level auditory features such

as pitch[28,29], and agree with the findings that event-related potential (ERP) responses to imagery loudness correlate with loudness responses to overt sounds[30]. Our results significantly extend those findings by demonstrating that auditory imagery can modulate neural responses to a most basic auditory attribute—loudness—and influence the perception of this perceptual attribute.

In addition to BE1, in which we observed the modulatory effect of auditory imagery on loudness perception, three additional behavioural experiments further substantiate the argument that the effects are caused by the low-level perceptual feature of loudness, by testing noise interference in BE2, accumulation over time in BE3 and the specificity of imagery loudness adaptation in the BE4. In particular, the results of BE4 show that similar modulation effects were observed even though the imagery content and subsequent auditory stimuli were different. These results demonstrate that the modulatory effects of imagery on loudness perception are not content specific. This is probably because loudness is a basic auditory attribute that is ubiquitously associated with auditory processing. Auditory imagery can reconstruct a similar representation of loudness to the one established during auditory perception, and hence can modulate loudness perception regardless of the particular representational content.

The MEG and EEG results consistently demonstrate that the modulation effects of imagery on loudness perception are reflected in the early auditory responses about 100 ms after stimulus onset. These results suggest that the top-down-induced reconstruction of representations without external stimulation can interact with the bottom-up perception in the early processing stream. Moreover, in a previous study[5], we found that imagery of an auditory syllable can modulate later auditory responses at 200 ms to an overt syllable sound. These differences in the temporal dynamics of modulation effects suggest that the interaction between the top-down-induced information and bottom-up perceptual process depends on the level of abstraction of the reconstructed perceptual attribute.

Our results may seem similar to a broader contextual effect on perception. For example, people judge the luminance of a grey patch to be higher on a black background than on a white background. The adaptation-level theory[31,32] and a later mathematical form as normalization[33,34] explain these contextual effects based on recalibration of the target by pooling or integrating the representation from the adjacent visual fields. This mechanism is different from ours in that the imagery-induced adaptation is caused by the repetitive activation of the same auditory representation. Moreover, attention and expectation could play a role in our results, as they are common factors in temporal cueing paradigms. A possible account of our observations is that the cue for imagined speech provides an anchoring bias for the loudness rating[32]. That is, seeing the visual display loud or soft will prime the judgement of loudness. This hypothesis is less likely to explain our findings, since we found the opposite pattern: participants rated a sound as more soft after they saw a visual cue in the loud condition and imagined speaking loudly. Furthermore, the MEG activity revealed an early (in time and in neural processing hierarchy) effect, typically interpreted as perceptual, rather than a later effect that could be interpreted as decisional.

Qualitatively, imagery-induced modulation effects on loudness perception are weaker than the effects caused by physical stimulation. Previous studies[35] found that if the intensity of preceding stimuli was significantly larger than the probe sound, or when they were at a similar intensity level, the overt auditory stimuli recalibrated the loudness perception to the subsequent stimuli at a smaller degree (~1–3 dB), whereas when the intensity of two stimuli differed within a modest range (for example, a preceding sound at 80 dB and a probe at 60–70 dB), the loudness adaptation could be as large as 10 dB. Imagery-induced modulation effects on loudness perception are about 0.25 in the behavioural rating and about 1 dB in neural responses. Assuming ratings in the current study are close to a linear scale, the imagery-induced adaptation is much weaker than the loudness recalibration caused by physical stimuli. Moreover, loudness recalibration by external stimulation reflects a change in the underlying representation of auditory intensity rather than a decisional criterion shift[36], similar to the results we obtained in the current study, which suggests that the imagery-induced modulation effects were caused by low-level perceptual processes during imagery.

The direction of imagery modulation effects has been observed as facilitation in the visual domain[17,37]. However, this study, as well other imagery studies in the auditory domain[38], found that auditory imagery shows suppression effects on perception. The apparent differences between facilitation and suppression may be caused by the different mechanisms in visual and auditory domains. Unlike visual processes, in which observers usually receive information from the external word, auditory processes—especially in the context of speech—are tightly linked to production. Extending this perception–production link in auditory processes to imagery, we can have two distinct types of speech imagery—imagined speaking and imagined hearing. In fact, in previous studies, we found that different types of speech imagery can either increase or decrease neural responses to auditory stimuli[5]. Moreover, the facilitation or suppression effects can be switched by the distance between the content of imagery and perception[39]. Therefore, we put forward the hypothesis that imagined speaking is mediated by motor-based mechanisms, whereas imagined hearing is underpinned by memory-based mechanisms[4]. Furthermore, we hypothesize that motor-based mechanisms would induce more precise representations than memory-based mechanisms, and such functional differences could be the cause of distinct direction in the modulation effects of speech imagery[5]. Preliminary functional magnetic resonance imaging evidence supports this dual-route mechanism[6].

Our results may implicate a possible mechanism of auditory hallucinations. A recent study[40] argues that auditory verbal imagery can cause highly hallucination-prone participants to be more willing to report hearing a voice in noise. In fact, neural evidence suggests that schizophrenic patients with auditory hallucination symptoms are more sensitive to internally induced sounds[41,42], and show decreased source and error monitoring functions[43,44]. Our results suggest that auditory imagery can induce detailed auditory representations, and such top-down-induced representations can interact with bottom-up process. We hypothesize that imagined speaking is mediated by a motor-to-sensory transformation mechanism, and this mechanism may be intact in patients with auditory hallucination to reconstruct representation internally, but the monitoring function that labels the source of origin may malfunction and attributes the internally induced presentation to external sources[4]. Therefore, the source of internally induced sounds could be confused with external sounds, resulting in more false positives in detection and auditory hallucination.

The fact that the 'loudness of thought' influences the loudness of hearing provides important evidence suggesting that top-down reconstructed neural representations converge to the same representational format as the bottom-up constructed representations in perception. Such representational overlap evidently extends to the processing of basic acoustic features, such as sound intensity. This adds to previous findings on imagery-induced modulation of perception[17,38], and suggests that such a coordinated transformation in a top-down process forms the neurocomputational foundation that enables the interaction with a bottom-up process. In addition, together with the findings that later auditory responses correlate with task demands on a more abstract (for example, syllabic) level[5], we demonstrate that the level of processing in the neural hierarchy of top-down-induced information depends on the demands of reconstructing perceptual attributes, which in turn constrains the interaction between top-down and bottom-up processes and the modulation effects on perception.

Methods

BE1: loudness judgement after imagined speech. BE1 tested how internally generated auditory imagery representations modulate loudness ratings for subsequent auditory stimuli.

Participants. A total of 16 undergraduate students (8 males; mean age: 20.6 years, range: 19–23 years) at New York University took part in this experiment for course credits. All participants were right handed. The experiment was approved by the Institutional Review Board (IRB) at New York University. Written informed consent was obtained from all participants.

Materials. Auditory stimuli were recorded in a sound-attenuated testing room using a Shure Beta 58 A microphone. Participants pronounced the syllable 'da' ten times. The auditory signals were recorded (sampling rate 44.1 kHz) and further processed using Praat (http://www.fon.hum.uva.nl/praat/). Participants wore Sennheiser HD 280 headphones when listening to the continuous recording and selected one auditory token as the stimulus. The mean duration of the selected auditory tokens was 325 ms. All sounds were normalized by average intensity (RMS) and 5 different loudness levels were then created (58, 59, 60, 61 or 62 dB SPL). The reason for choosing the 1 dB step size—approximately 1 just-noticeable-difference (JND) for loudness—was that the effect size of imagery could be small. If the loudness step size was too large, it could miss small modulation effects caused by the imagery. After the basic phenomenon was established, we increased the resolution to 2 dB steps for the rest of experiments.

Procedure. Before the experiment, participants were familiarized with the five levels of intensity by listening to them in a loop with increasing loudness three times. After confirming the perceptual distinction of different levels of loudness, participants performed the experiment. On each trial, participants were asked to imagine speaking the syllable 'da' at two different loudness levels (soft versus loud), followed by making a loudness judgement about the subsequent auditory stimuli (Fig. 1a). A cue word—'loud' or 'soft'—randomly appeared and stayed in the centre of the screen until a button press. This cue word informed participants whether they should imagine using the loudest possible voice (loud) or the softest possible voice (soft). After a blank 500 ms interval following the offset of the cue word, the written syllable 'da' flashed four times in a row, each with a duration of 500 ms, and each with a 500 ms blank in between, thus the stimulus onset asynchrony was 1,000 ms. Participants were asked to imagine speaking the syllable 'da' four times, synchronized with the onset of each visual presentation at a given loudness level indicated by the initial cue word. After the imagined speech, a fixation cross appeared at the centre of the screen, with the duration jittered between 750 ms and 1,250 ms (125 ms increments). An overt 'da' syllable was presented at a random intensity level (levels 1–5, corresponding to 58–62 dB SPL) after the offset of fixation. Participants were asked to judge the loudness level by pressing 1 to 5, where 1 was softest and 5 was loudest.

The factor imagined loudness (two levels: soft versus loud) was fully crossed with the factor sound intensity (five levels) to yield ten conditions. Four blocks were included in this experiment, each with 50 trials (5 trials per condition in each block; 20 trials per condition in total) presented in random order within a block. A microphone was placed next to the participant's mouth to monitor and confirm that there was no overt pronunciation throughout the experiment.

Data analysis. The loudness judgements were analysed using a repeated-measures two-way ANOVA (with the factors imagined loudness and sound intensity).

BE2: loudness judgement after imagined speech in noise. Inspired by an imagery study in the visual domain[17], we designed three more behavioural experiments to further verify that the modulation effects of imagery obtained in BE1 reflected low-level perceptual adaptation. BE2 examined whether the modulatory effects of imagery could be abolished by presenting noise in the period when the participants imagined speaking (Fig. 1a). The rationale was that although imagined speech can possibly modify auditory representations, physically presented noise has a more direct influence on auditory representations that could override the speech imagery effect.

Participants. A total of 16 students (6 males; mean age: 22.3 years, range: 19–25 years) at East China Normal University took part in this experiment for monetary compensation. All participants were right handed. The experiment was approved by the IRB at New York University Shanghai. Written informed consent was obtained from all participants.

Materials. Auditory stimuli were the same as in BE1, except that 5 different loudness levels with a larger spacing were used (58, 60, 62, 64 and 66 dB SPL).

Procedure. The experimental procedure was the same as in BE1, except that acoustic white noise was delivered at about 60 dB SPL for 4 s during the imagery period, starting from the onset of the first visual display of 'da' and ending at the onset of the fixation.

Data analysis. The loudness judgements were analysed using a repeated-measures two-way ANOVA (with the factors imagined loudness and sound intensity).

BE3: loudness judgement after different repetitions of imagined speech. This experiment further tested whether the observed effects in BE1 were perceptual in nature, by examining whether the strength of modulatory effects of imagery can be increased by imagining speaking a syllable loudly or softly more times. Moreover, the visual display of the syllable 'da' was replaced with a fixation to avoid the potential influence of reading.

Participants. A total of 16 students (5 males; mean age: 23.1 years, range: 20–26 years) at East China Normal University took part in this experiment for monetary compensation. All participants were right handed and did not participate in BE1 or BE2. The experiment was approved by the IRB at New York University Shanghai. Written informed consent was obtained from all participants.

Materials. Auditory stimuli (the syllable 'da') were self-recorded for each individual participant and the standard was the same as in BE2.

Procedure. The experimental procedure was similar to that of BE1, with two exceptions (Fig. 2a). First, the visual display of the syllable 'da' was replaced by a fixation cross. Participants were required to imagine speaking the syllable 'da' synchronized with the onset of each visual presentation of fixation at a given loudness level indicated by the initial cue word. The use of fixation avoided the possible influence of reading. Second, the number of times fixation appeared in each trial was randomized between one and six. Therefore, the speech imagery was repeated a different number of times for different trials. After the imagined speech, a red fixation cross appeared at the centre of the screen, with its duration jittered between 750 ms and 1,250 ms (125 ms increments). This red fixation indicated the end of the imagery period and was followed by an overt 'da' syllable at a random intensity level (levels 1–5, corresponding to 58–66 dB SPL) for which participants were asked to make a loudness judgement. Three factors, imagined loudness (2 levels: soft versus loud), sound intensity (5 levels) and imagery repetitions (6 levels) were fully crossed and yielded 60 conditions. Four blocks were included in this experiment, each with 60 trials (1 trial per condition in each block; 4 trials per condition in total) presented in a random order.

Data analysis. The difference in rating scores between the loud and soft conditions was analysed, along with how this measure was influenced by the number of repetitions of speech imagery. To increase the statistical power, the rating difference was averaged across five intensity levels. The number of repetitions for the speech imagery was divided into 3 bins (1 or 2, 3 or 4, or 5 or 6 repetitions) yielding 3 levels. The loudness judgements rating difference was analysed by a repeated-measures one-way ANOVA, the factor being the imagery repetitions (3 levels).

BE4: loudness judgement after imagined counting. This experiment further extended the procedure of BE3 by changing to a covert counting number task. The purpose of this experiment was threefold: first, to replicate the results obtained in BE3. Second, to verify that the participants indeed imagined speaking in the experiment. Third, to examine the content specificity of imagery effects in the previous behavioural experiments (that is, whether the loudness rating effect is specific to the imagined content).

Participants. A total of 16 students (4 males; mean age: 23.8 years, range: 19–27 years) at East China Normal University took part in this experiment for monetary compensation. All participants were right handed and did not participate in the other three behavioural experiments. The experiment was approved by the IRB at New York University Shanghai. Written informed consent was obtained from all participants.

Materials. Auditory stimuli (the syllable 'da') were self-recorded for each individual participant and the standard was the same as in BE3.

Procedure. The experimental procedure was similar as in BE3, with two exceptions. First, participants were required to covertly count the number of fixations; for example, one, two, three (in Chinese), synchronized with the onset of each visually presented fixation at the given loudness level indicated by the initial cue word. Second, a dual-task design was employed. After the overt 'da' syllable at a random intensity level (levels 1–5, corresponding to 58–66 dB SPL), the participants were asked to perform one of two possible tasks indicated by the visual instruction. They were instructed to perform the loudness judgement task on 83.3% trials. In the other 16.7% trials, participants performed a counting judgement task for which they had to judge whether a number shown on the screen was the same as the number they counted (that is, the number of fixations shown in that trial). The number shown on the screen either matched the number they counted or was bigger by 1.

For the loudness judgement trials, the same three factors, imagined loudness (2 levels: soft versus loud), sound intensity (5 levels) and imagery repetitions

(6 levels) were fully crossed and yielded 60 conditions. In a block, 60 trials of loudness judgement task were included (one trial per condition in each block; four trials per condition in total). For the trials of the counting judgement task, 12 trials were included in each block, with 1 trial for each of the 12 conditions (2 levels of imagery loudness × 6 levels of imagery repetitions), but randomly selected out of the 5 levels of sound intensity. Four blocks were included in this experiment with 72 trials in each block. All types of trial were presented in a random order within a block.

Data analysis. The average accuracy of the counting judgement task was obtained by averaging all trials in this task. For the loudness judgement task, trials were first averaged across 5 levels of sound intensity, and further averaged across every second of imagery repetitions (1 or 2, 3 or 4, or 5 or 6) yielding 3 levels of imagery repetitions. The difference rating scores between loud and soft were calculated (loud minus soft) and the loudness judgements rating differences were analysed using a repeated-measures one-way ANOVA, the factor being imagery repetitions (three levels).

MEG experiment: imagined speech induces loudness neural adaptation. The MEG experiment was conducted to test how the auditory neural responses were modulated by the preceding internally generated sound (loud versus soft).

Participants. A total of 16 participants (12 males; mean age: 30.6 years, range: 22–55 years) took part in this experiment, for pay. None of these participants was included in any of the behavioural experiments. All participants were right handed without any history of neurological disorders. This experiment was approved by the New York University IRB. Written informed consent was obtained from all participants.

Materials. Auditory stimuli were recorded in a sound-attenuated room using a Radio Shack Unidirectional Dynamic 33-3002 microphone. Participants pronounced the syllable 'da' ten times using their normal, most comfortable pitch. The continuous auditory signals were recorded (at a sampling rate of 44.1 kHz) and further processed using Praat. Participants listened to the continuous recording and selected one auditory token as a stimulus. The mean duration of the selected auditory tokens was 301 ms. All sounds were normalized by average intensity (RMS) to 70 dB SPL and delivered through plastic air tubes connected to foam ear pieces (E-A-R Tone Gold 3A Insert earphones; Aearo Technologies).

Procedure. The MEG experimental procedure was similar to that of BE1, except for several modifications optimized for the electrophysiological recordings. First, a no-imagery condition was included in the MEG experiment to establish a baseline condition for the soft and loud conditions. During the no-imagery condition, participants were not asked to imagine speaking; instead, they passively saw the symbols '##' being flashed four times. Second, only one sound intensity level was used. Participants were asked to passively listen to the sound and were not required to make a loudness judgement. Therefore, only three conditions were included in this MEG experiment (soft, loud and no-imagery; the factor being imagined loudness). We set a microphone next to the participants to check that there was no overt pronunciation throughout the experiment. Four blocks were included in the experiment, with 45 trials in each block (15 trials per condition in each block; 60 trials per condition in total). The stimulus presentation order was randomized.

MEG recording. Neuromagnetic signals were measured using a 157-channel whole-head axial gradiometer system (Kanazawa Institute of Technology). Five electromagnetic coils were attached to each participant's head to monitor their 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 three-dimensional digitizer software (Source Signal Imaging) and digitizing hardware (Polhemus). 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 and 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 principal component analysis method[45]. Trials with amplitudes >3 pT (~5%) were considered artefacts and discarded. For each condition, epochs of response to the auditory probe, 400 ms in duration including a 100 ms pre-stimulus period, were extracted. The averages were low-pass filtered with a cutoff frequency of 30 Hz. A typical M100/M200 auditory response complex was observed[24] and the peak latencies were identified for each individual participant.

Because of possible interactions between the neural response magnitude and neural source distribution at the sensor level[22,23], a multivariate measurement technique ('angle test of response similarity') was implemented to assess the topographical similarity between the auditory responses in the three conditions (loud, soft and no imagery). This technique allows the assessment of spatial similarity in electrophysiological studies regardless of the response magnitude, and estimates the similarities in the distribution of underlying neural sources (for example, refs [5,39,46–51]). Using 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 two topographies. The cosine value of this angle, which is called the angle measure, can be calculated from the dot product of these two response vectors where value 1 stands for an exact match (angle equals zero) and the value −1 stands for the opposite (angle equals π).

The angle measure between topographies (that is, the between angle measure) in different conditions is statistically tested against a null hypothesis (that is, the angle between two topographical patterns greater than chance). The null hypothesis is formed by comparing the pattern similarity of responses after randomly shuffling among sensors (shuffled angle measure). In this study, the between angle measure was calculated between auditory responses in pairs among three conditions (for example, soft versus no imagery, loud versus no imagery, and soft versus loud). The null distribution was formed by repeatedly computing the angle measures in the same pairs 10,000 times, but using the shuffled topographies, and the maximum value in this distribution was selected as the shuffled angle measure. The between angle measure was compared with the shuffled angle measure of a given pair (using paired *t*-tests) to statistically determine the topographical similarity between conditions. If the between angle measure is significantly smaller than the shuffled angle measure (that is, the angle between two topographies is greater than chance), the two topographies are different and hence we infer distinct neural source distributions. In contrast, if the between angle measure is significantly larger than the shuffled angle measure, it suggest the two topographies are similar and the following magnitude test could be free of confounds of source distribution changes.

After confirming the stability of the neural source distributions among conditions, any observed changes in the sensor level analysis are attributed to the response magnitude change. The RMS of waveforms across 157 channels, indicating the global response power, was calculated and employed in the following statistical tests. A 25 ms time window centred 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 in the no-imagery condition from the one in the soft and loud conditions. Paired *t*-tests were carried out between the relative changes between the soft and loud conditions for the M100 and M200 responses.

Distributed source localization of the adaptation effects was obtained using minimum-norm estimation software (https://martinos.org/mne/stable/index.html; Martinos Center for Biomedical Imaging, Massachusetts General Hospital). L2 minimum-norm current estimates were constrained on the cortical surface that was reconstructed from individual structural magnetic resonance imaging data with FreeSurfer software (https://surfer.nmr.mgh.harvard.edu/; Martinos Center for Biomedical Imaging, Massachusetts General Hospital). Current sources were about 5 mm apart on the cortical surface, yielding approximately 2,500 locations per hemisphere. Because MEG is sensitive to electromagnetic fields generated from current sources that are in sulci and tangential to the cortical surface[52], deeper sources are given more weight to overcome the minimum-norm estimation bias towards superficial currents, and current estimation favours the sources normal to the local cortical surface[53]. Individual single-compartment boundary element models were used to compute the forward solution. Based on the forward solution, the inverse solution was calculated by approximating the current source spatiotemporal distribution that best explains the variance in the observed MEG data. To compute and visualize the minimum-norm estimation group results, each participant's cortical surface was inflated and flattened[54] and morphed to a representative surface[55]. Current estimation was first performed within each condition and the adaptation effects were obtained by subtracting the absolute values of estimation in the loud condition from those in the soft condition and then averaged across participants. The same M100 and M200 time windows as those used in event-related analysis were then applied. A common measure of the normalized response magnitude, dynamic statistical parametric mapping, was obtained for each condition[56]. Essentially, it is the source strength at the time and space of interest, normalized by the variance during a baseline period. We used this measure to indicate the anatomical location of the modulation effects that we found in source space, by taking the differences of dynamic statistical parametric mapping values between the soft and loud conditions.

EEG experiment: imagined speech induces loudness neural adaptation, which correlates with behaviour. The EEG experiment was designed to replicate the results of the MEG experiment in addition to investigating the relationship between neural adaptation and behavioural ratings of loudness.

Participants. A total of 18 participants (7 males; mean age: 23.4 years, range: 20–29 years) participated in this experiment, for pay. None of the participants was included in the behavioural or MEG experiments. All participants were right handed without any history of neurological disorders. This experiment was approved by the New York University Shanghai IRB. Written informed consent was obtained from all participants.

Materials. Auditory stimuli were recorded in a sound-attenuated room using a Shure Beta 58A microphone. Participants pronounced the syllable 'da' ten times using their normal, most comfortable pitch. The continuous auditory signals were recorded (at a sampling rate of 44.1 kHz) and further processed using Praat. Participants listened to the continuous recording and selected one auditory token as a stimulus. The mean duration of the selected auditory tokens was 339 ms. All sounds were normalized by average intensity (RMS) and five different loudness levels were created (58, 60, 62, 64 and 66 dB SPL) and delivered through plastic air tubes connected to foam ear pieces (ER-3C Insert Earphones; Etymotic Research).

Procedure. The EEG experimental procedure was similar to that used in the MEG experiment, except for several modifications optimized for investigating the relationship between neural adaptation and behavioural performance. First, 5 sound intensity levels were used, with a space of 2 dB between adjacent levels to increase perceptual differences. Second, participants were required to rate the loudness of the auditory stimuli by pressing 1–5, where 1 was softest and 5 was loudest. Therefore, 15 conditions were included in this EEG experiment (the factor 'imagined loudness': soft, loud and no imagery; and the factor 'sound intensity': 5 levels). We set a microphone next to participants to check that there was no overt pronunciation throughout the experiment. Four blocks were included in the experiment, with 45 trials in each block (15 trials per condition in each block; 60 trials per condition in total). The stimulus presentation order was randomized.

EEG recording. EEG signals were measured using a 32-channel active electrode system (Brain Vision actiCHamp; Brain Products). Electrodes were placed on an ActiCap, on which electrode holders were arranged according to the 10–20 international electrode system. The impedance of each electrode was kept below 5 kΩ and the data were referenced online to the electrode of Cz and re-referenced offline to the average of all electrodes. The EEG data were acquired with Brain Vision PyCoder software (http://www.brainvision.com/pycorder.html) with a sampling frequency of 1,000 Hz and filtered online between DC and 200 Hz, with a notch filter at 50 Hz.

EEG analysis. EEG signal processing and analysis were carried out in MATLAB using the EEGlab[57] and ERPlab toolboxes[58]. For each condition, epochs of response to the auditory probe, 400 ms in duration including a 100 ms pre-stimulus period and a 300 ms post-stimulus period, were extracted. The averages were low-pass filtered with a cutoff frequency of 30 Hz. A typical N1/P2 auditory response complex was observed and the peak latencies were identified for each individual participant.

The RMS of waveforms across 32 electrodes, indicating the global response power, was calculated separately for three imagined loudness conditions (no imagery, loud and soft) and employed in the following statistical tests. A 25 ms time window centred at individual N1 and P2 latencies was applied to obtain the temporal average responses. The relative response power changes were further calculated by subtracting the responses in the no-imagery condition from those in the soft and loud conditions, and then the relative response changes were normalized by dividing the responses in the no-imagery condition. Paired *t*-tests were carried out between the normalized relative changes in the soft and loud conditions for the N1 and P2 responses.

To investigate the relationship between neural adaptation and behavioural ratings of loudness, a median split of each condition was carried out. Specifically, trials of 3 middle-intensity levels (60, 62 and 64 dB, to avoid ceiling or flooring effects) in the loud and soft conditions were ranked by the loudness rating scores from low to high and separated into two groups—a 'rating lower' group and a 'rating higher' group—that contained equal numbers of trials. RMS of waveforms across 32 electrodes and temporal averaged responses at N1 and P2 latencies were obtained separately for rating higher and rating lower. Paired *t*-tests were carried out between rating higher and rating lower for the N1 and P2 responses.

For the correlation analysis between the behavioural and neural responses, trials at the 3 middle-intensity levels (60, 62 and 64 dB) were included, to avoid ceiling and floor effects. Behavioural response decreases were obtained by subtracting ratings in the no-imagery condition from those in the imagery conditions (loud and soft). Neural response (N1) differences were quantified by converting the N1 response magnitude differences into neurometric dB form using the following calculation: $dB = 10*\log_{10}(N1_{imagery}) - 10*\log_{10}(N1_{no\text{-}imagery})$. Pearson correlation analysis was conducted between the behavioural and neural response differences. A linear regression analysis was carried out using the same set of data.

To rule out the possibility that the observed modulation effects on the responses to the auditory stimuli may be caused by the lingering neural activity induced by the preceding imagery (although this is less likely since the imagery activity was short-lived and there was an average of 1 s gap between the end of imagery and the onset of auditory stimuli), we examined whether the neural activity in the baseline period (100 ms before the onset of auditory stimuli) was different between the imagery (loud versus soft) and no-imagery conditions. Because this analysis was to test the null hypothesis, we used a Bayesian analysis method for paired *t*-tests[59] (online tool at http://pcl.missouri.edu/bf-one-sample). The Bayes factor B01 = M0/M1, where M0 and M1 are the marginal likelihood for the null and alternative, respectively. That is, the Bayes factors are odds ratios between the null and alterative hypothesis, which means that the null is B01 times more probable than the alternative. The parameters we input for the Bayesian analysis were a sample size of 18 and scale *r* on an effect size of 0.707. For the evoked responses, the temporal averages in the time window of 100 ms before the stimulus were obtained for each condition. For the induced responses, activity was obtained for the same baseline period before the stimulus and separate periods for the alpha (9–12 Hz), beta (13–25 Hz) and low gamma (26–41 Hz) bands, because the period after the last imagery task and before the auditory probe was not long enough (shortest of 750 ms) to reliably estimate activity in the delta (1–2 Hz) and theta (3–8 Hz) bands. Both evoked and induced responses were subject to *t*-tests for the pair of loud and no imagery, the pair of soft and no imagery, the pair of loud and soft, and the pair of no imagery and the average of loud and soft. The Bayes factor was obtained for *t*-values of each comparison.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The custom code used in this study is available from the corresponding author upon reasonable request.

Data availability. The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Received: 18 January 2017; Accepted: 16 January 2018;
Published online: 19 February 2018

References

1. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).
2. Firestone, C. & Scholl, B. J. Cognition does not affect perception: evaluating the evidence for 'top-down' effects. *Behav. Brain Sci.* **39**, e229 (2016).
3. Molloy, K., Griffiths, T. D., Chait, M. & Lavie, N. Inattentional deafness: visual load leads to time-specific suppression of auditory evoked responses. *J. Neurosci.* **35**, 16046–16054 (2015).
4. Tian, X. & Poeppel, D. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Front. Hum. Neurosci.* **6**, 314 (2012).
5. Tian, X. & Poeppel, D. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *J. Cogn. Neurosci.* **25**, 1020–1036 (2013).
6. Tian, X., Zarate, J. M. & Poeppel, D. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex* **77**, 1–12 (2016).
7. Wheeler, M. E., Petersen, S. E. & Buckner, R. L. Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proc. Natl Acad. Sci. USA* **97**, 11125–11129 (2000).
8. Kosslyn, S. M., Ganis, G. & Thompson, W. L. Neural foundations of imagery. *Nat. Rev. Neurosci.* **2**, 635–642 (2001).
9. Zatorre, R. J. & Halpern, A. R. Mental concerts: musical imagery and auditory cortex. *Neuron* **47**, 9–12 (2005).
10. Kosslyn, S. M. et al. The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science* **284**, 167–170 (1999).
11. Slotnick, S. D., Thompson, W. L. & Kosslyn, S. M. Visual mental imagery induces retinotopically organized activation of early visual areas. *Cereb. Cortex* **15**, 1570–1583 (2005).
12. Thirion, B. et al. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* **33**, 1104–1116 (2006).
13. Bunzeck, N., Wuestenberg, T., Lutz, K., Heinze, H.-J. & Jancke, L. Scanning silence: mental imagery of complex sounds. *Neuroimage* **26**, 1119–1127 (2005).
14. Halpern, A. R. & Zatorre, R. J. When that tune runs through your head: a PET investigation of auditory imagery for familiar melodies. *Cereb. Cortex* **9**, 697–704 (1999).
15. Kosslyn, S. M. & Thompson, W. L. When is early visual cortex activated during visual mental imagery? *Psychol. Bull.* **129**, 723–746 (2003).
16. Tartaglia, E. M., Bamert, L., Mast, F. W. & Herzog, M. H. Human perceptual learning by mental imagery. *Curr. Biol.* **19**, 2081–2085 (2009).
17. Pearson, J., Clifford, C. W. & Tong, F. The functional impact of mental imagery on conscious perception. *Curr. Biol.* **18**, 982–986 (2008).
18. Pearson, J., Rademaker, R. L. & Tong, F. Evaluating the mind's eye: the metacognition of visual imagery. *Psychol. Sci.* **22**, 1535–1542 (2011).
19. Laeng, B. & Sulutvedt, U. The eye pupil adjusts to imaginary light. *Psychol. Sci.* **25**, 188–197 (2014).
20. Scott, M. Corollary discharge provides the sensory content of inner speech. *Psychol. Sci.* **24**, 1824–1830 (2013).
21. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).
22. Tian, X. & Huber, D. E. Measures of spatial similarity and response magnitude in MEG and scalp EEG. *Brain Topogr.* **20**, 131–141 (2008).

23. Tian, X., Poeppel, D. & Huber, D. E. TopoToolbox: using sensor topography to calculate psychologically meaningful measures from event-related EEG/MEG. *Comput. Intell. Neurosci.* **2011**, 674605 (2011).
24. Roberts, T. P. L., Ferrari, P., Stufflebeam, S. M. & Poeppel, D. Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception. *J. Clin. Neurophysiol.* **17**, 114–129 (2000).
25. Kraemer, D. J., Macrae, C. N., Green, A. E. & Kelley, W. M. Musical imagery: sound of silence activates auditory cortex. *Nature* **434**, 158 (2005).
26. Oh, J., Kwon, J. H., Yang, P. S. & Jeong, J. Auditory imagery modulates frequency-specific areas in the human auditory cortex. *J. Cogn. Neurosci.* **25**, 175–187 (2013).
27. Linke, A. C. & Cusack, R. Flexible information coding in human auditory cortex during perception, imagery, and STM of complex sounds. *J. Cogn. Neurosci.* **27**, 1322–1333 (2015).
28. Cebrian, A. N. & Janata, P. Electrophysiological correlates of accurate mental image formation in auditory perception and imagery tasks. *Brain Res.* **1342**, 39–54 (2010).
29. Herholz, S. C., Lappe, C., Knief, A. & Pantev, C. Neural basis of music imagery and the effect of musical expertise. *Eur. J. Neurosci.* **28**, 2352–2360 (2008).
30. Wu, J., Yu, Z., Mai, X., Wei, J. & Luo, Y. Pitch and loudness information encoded in auditory imagery as revealed by event-related potentials. *Psychophysiology* **48**, 415–419 (2011).
31. Helson, H. Current trends and issues in adaptation-level theory. *Am. Psychol.* **19**, 26–38 (1964).
32. Stevens, S. S. Adaptation-level vs. the relativity of judgment. *Am. J. Psychol.* **71**, 633–646 (1958).
33. Heeger, D. J. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
34. Louie, K., Grattan, L. E. & Glimcher, P. W. Reward value-based gain control: divisive normalization in parietal cortex. *J. Neurosci.* **31**, 10627–10639 (2011).
35. Mapes-Riordan, D. & Yost, W. A. Loudness recalibration as a function of level. *J. Acoust. Soc. Am.* **106**, 3506–3511 (1999).
36. Arieh, Y. & Marks, L. E. Recalibrating the auditory system: a speed-accuracy analysis of intensity perception. *J. Exp. Psychol. Hum. Percept. Perform.* **29**, 523–536 (2003).
37. Brascamp, J. W., Knapen, T. H., Kanai, R., van Ee, R. & van den Berg, A. V. Flash suppression and flash facilitation in binocular rivalry. *J. Vision.* **7**, 12.1-12 (2007).
38. Scott, M. Corollary discharge provides the sensory content of inner speech. *Psychol. Sci.* **24**, 1824–1830 (2013).
39. Tian, X. & Poeppel, D. Dynamics of self-monitoring and error detection in speech production: evidence from mental imagery and MEG. *J. Cogn. Neurosci.* **27**, 352–364 (2015).
40. Moseley, P., Smailes, D., Ellison, A. & Fernyhough, C. The effect of auditory verbal imagery on signal detection in hallucination-prone individuals. *Cognition* **146**, 206–216 (2016).
41. Ford, J. M. et al. Neurophysiological studies of auditory verbal hallucinations. *Schizophr. Bull.* **38**, 715–723 (2012).
42. Ford, J. M. et al. Tuning in to the voices: a multisite FMRI study of auditory hallucinations. *Schizophr. Bull.* **35**, 58–66 (2009).
43. Mathalon, D. H., Jorgensen, K. W., Roach, B. J. & Ford, J. M. Error detection failures in schizophrenia: ERPs and fMRI. *Int. J. Psychophysiol.* **73**, 109–117 (2009).
44. Perez, V. B. et al. Error monitoring dysfunction across the illness course of schizophrenia. *J. Abnorm. Psychol.* **121**, 372–387 (2012).
45. De Cheveigné, A. & Simon, J. Z. Denoising based on time-shift PCA. *J. Neurosci. Methods* **165**, 297–305 (2007).
46. Almeida, D. & Poeppel, D. Word-specific repetition effects revealed by MEG and the implications for lexical access. *Brain Lang.* **127**, 497–509 (2013).
47. Davelaar, E. J., Tian, X., Weidemann, C. T. & Huber, D. E. A habituation account of change detection in same/different judgments. *Cogn. Affect. Behav. Neurosci.* **11**, 608–626 (2011).
48. Huber, D. E., Tian, X., Curran, T., O'Reilly, R. C. & Woroch, B. The dynamics of integration and separation: ERP, MEG, and neural network studies of immediate repetition effects. *J. Exp. Psychol. Hum. Percept. Perform.* **34**, 1389–1416 (2008).
49. Luo, H., Tian, X., Song, K., Zhou, K. & Poeppel, D. Neural response phase tracks how listeners learn new acoustic representations. *Curr. Biol.* **23**, 968–974 (2013).
50. Tian, X. & Huber, D. E. Playing "duck duck goose" with neurons: change detection through connectivity reduction. *Psychol. Sci.* **24**, 819–827 (2013).
51. Tian, X. & Poeppel, D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* **1**, 166 (2010).
52. Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J. & Lounasmaa, O. V. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Rev. Mod. Phys.* **65**, 413–497 (1993).
53. Lin, F. H., Belliveau, J. W., Dale, A. M. & Hämäläinen, M. S. Distributed current estimates using cortical orientation constraints. *Hum. Brain Mapp.* **27**, 1–13 (2006).
54. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9**, 195–207 (1999).
55. Fischl, B., Sereno, M. I., Tootell, R. B. H. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).
56. Dale, A. M. et al. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* **26**, 55–67 (2000).
57. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
58. Lopez-Calderon, J. & Luck, S. J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front. Hum. Neurosci.* **8**, 213 (2014).
59. Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D. & Iverson, G. Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16**, 225–237 (2009).

Acknowledgements

We thank J. Walker for technical support with MEG data collection, S. Yuan for help with running the EEG experiment, Q. Xu for help with running BE2–BE4 and L. Tao for comments and edits on an early draft. This study was supported by the National Natural Science Foundation of China (31500914 to X.T., and 31771248 and 31500873 to N.D.), the Major Program of the Science and Technology Commission of Shanghai Municipality (15JC1400104 and 17JC1404104), the Program of Introducing Talents of Discipline to Universities (Base B16018), a grant from the New York University Global Seed Grants for Collaborative Research (85-65701-G0757-R4551), the Joint Research Institute Seed Grants for Research Collaboration from the New York University-East China Normal University Institute of Brain and Cognitive Science at New York University, Shanghai (to X.T.), the Zhejiang Provincial Natural Science Foundation of China (LR16C090002), research funding from the State Key Laboratory of Industrial Control Technology, Zhejiang University (to N.D.) and National Institutes of Health 2R01DC05660 (to D.P.). No funders had any role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

X.Tian conceived and designed the study, performed BE1 and the MEG experiments and analysed the data. F.B. performed BE2–BE4 and the EEG experiments. X.Tian, N.D., X.Teng and D.P. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0305-8.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to X.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Xing Tian

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Experimental design

1. Sample size

Describe how sample size was determined.

The sample size was chosen based on previous experiments on speech imagery. The sample size for previous experiments is usually between 12 and 20 (e.g. Journal of Cognitive Neuroscience, 2013; 2015).

2. Data exclusions

Describe any data exclusions.

NA

3. Replication

Describe the measures taken to verify the reproducibility of the experimental findings.

Our findings are in line with large number of previous studies. We show consistent results in 4 related behavioral and 1 MEG and 1 EEG experiments.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

NA

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

NA

Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ Test values indicating whether an effect is present *Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.*
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Praat (Boersma, P. and D. Weenink, Praat (Version 6.0. 14)[Software]. Latest version available for download from www. praat. org, 2016.) for sound stimuli recording and manipulation. MNE toolbox for MEG source localization, EEGlab (Delorme, A. and S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods, 2004. 134(1): p. 9-21.) and ERPlab (Lopez-Calderon, J. and S.J. Luck, ERPLAB: an open-source toolbox for the analysis of event-related potentials. Frontiers in human neuroscience, 2014. 8: p. 213.) for EEG signals pre-processing. Matlab R2014a

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.	NA

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	NA

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	NA
b. Describe the method of cell line authentication used.	NA
c. Report whether the cell lines were tested for mycoplasma contamination.	NA
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	NA

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide all relevant details on animals and/or animal-derived materials used in the study.	no animal were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Sixty-four healthy volunteers in 4 behvioarl experiments: average age: 22.5 (range 19 - 27), 23 males. Sixteen healthy volunteers in the MEG experiemnt: average age: 30.6 (range 22 - 55), 12 males. Eighteen healthy volunteers in the EEG experiment: average age: 23.4 (range 20 -29), 7 males. All participants are right-handed.