nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-00962-0

Check for updates

Speaking rhythmically can shape hearing

M. Florencia Assaneo[1,2,7]✉, Johanna M. Rimmele[3,7]✉, Yonatan Sanz Perl[4,5,6] and David Poeppel[1,3]

Evidence suggests that temporal predictions arising from the motor system can enhance auditory perception. However, in speech perception, we lack evidence of perception being modulated by production. Here we show a behavioural protocol that captures the existence of such auditory–motor interactions. Participants performed a syllable discrimination task immediately after producing periodic syllable sequences. Two speech rates were explored: a 'natural' (individually preferred) and a fixed 'non-natural' (2 Hz) rate. Using a decoding approach, we show that perceptual performance is modulated by the stimulus phase determined by a participant's own motor rhythm. Remarkably, for 'natural' and 'non-natural' rates, this finding is restricted to a subgroup of the population with quantifiable auditory–motor coupling. The observed pattern is compatible with a neural model assuming a bidirectional interaction of auditory and speech motor cortices. Crucially, the model matches the experimental results only if it incorporates individual differences in the strength of the auditory–motor connection.

Auditory perception, especially the analysis of speech and music, requires precise timing at a timescale of tens to hundreds of milliseconds. Several distinct theories posit that temporal predictions from the motor system, which operates at these timescales[1,2], can optimize auditory perception[3,4]. For example, the 'active sensing' framework suggests that the efferent motor signals that are generated when producing a sound are activated and used to predict the sensory input[2,5,6]. Supporting this line of argumentation, it has been shown that during passive listening—especially under adverse conditions[7–10]—motor cortex is recruited[11–13].

What mechanisms could underpin the relevant sensorimotor interactions? Neuronal oscillations are ubiquitous across cortex and have been observed in auditory cortex as well as motor areas[14,15]. Moreover, the entrainment of neuronal oscillations has been ascribed a crucial role in aligning neuronal excitability to the temporal dynamics of behaviourally relevant (quasi-periodic) events[14,16,17]. In line with these findings, it has been recently proposed that top-down effects from the motor system can phase-reset neuronal oscillations in auditory cortex to optimize perception, by aligning the neuronal excitability phase to the occurrence of an expected event[18,19]. We would expect the impact of the motor system on auditory oscillatory activity to be particularly relevant during speech perception, where production and perception tightly interact[9,20] and wherein the perceived acoustic signal possesses a quasi-rhythmic temporal structure[21,22]. Indeed, neurophysiological studies suggest that brain rhythms in the motor and auditory cortices are coupled during speech comprehension[23–25]. The implicitly assumed and widely discussed hypothesis that auditory–motor coupling observed during passive listening actually reflects temporal predictions from the motor system implies that speech perception can be shaped by speech production. However, the basic link remains unexplored.

The entrainment of neuronal oscillations to periodic stimulation can be investigated by showing resonance phenomena. That is, the alignment of neuronal excitability to a periodic stimulation results in behavioural performance benefits in certain phases of the stimulation compared with others, even after the stimulation ceases. While behavioural and neuronal resonance phenomena have been reported[26–31], conclusive evidence regarding auditory (and particularly speech) entrainment is sparse. As entrainment phenomena are subtle and overlaid by other processes, one possibility is that individual differences in entrainment complicate the investigation of the phenomenon. For example, it has been shown that individuals differ in their ability to align their tapping to rhythmic auditory stimulation[32] and, in the context of speech, that only a subgroup of the population is spontaneously compelled to synchronize the produced syllable rate to the perceived one[33].

The current work aims to shed light on the neural mechanisms underlying sensorimotor interaction. We aim to find behavioural evidence for the proposed auditory–motor entrainment during speech perception and to build a neural model compatible with the experimental outcome. Crucially, we integrate individual differences—discussed above—into our study. We capitalize on a test developed by our group[33] to quantify—and control for—individual differences. This test (the spontaneous speech synchronization test, SSS-test) reliably classifies participants as high or low synchronizers ('highs' or 'lows') according to the degree of observed auditory-to-motor speech synchronization, that is, the synchronization of participants' speech production to heard speech. Here, we hypothesize that these individual differences (high versus low synchronizers) could also be reflected in the other direction, that is, motor to auditory. Crucially, the hypothesis is grounded in previously reported neural data. On the one hand, highs showed enhanced structural connectivity between frontal and auditory regions and stronger brain-to-stimulus coupling during passive speech listening in the left inferior frontal gyrus (IFG)[33]. On the other hand, top-down signals from IFG to auditory areas have been shown to enhance speech perception[25,34].

We designed a behavioural protocol (Fig. 1a) to explore the entrainment of speech perception by periodic speech production, with entrainment facilitating syllable detection at certain phases of the self-generated motor rhythm (motor-to-auditory entrainment).

[1]Department of Psychology, New York University, New York, NY, USA. [2]Instituto de Neurobiología, Universidad Nacional Autónoma de México, Santiago de Querétaro, Mexico. [3]Department of Neuroscience, Max-Planck-Institute for Empirical Aesthetics, Frankfurt am Main, Germany. [4]Department of Physics, FCEyN, University of Buenos Aires, Buenos Aires, Argentina. [5]National Scientific and Technical Research Council (CONICET), Buenos Aires, Argentina. [6]University of San Andrés, Buenos Aires, Argentina. [7]These authors contributed equally: M. Florencia Assaneo, Johanna M. Rimmele. ✉e-mail: fassaneo@gmail.com; johanna.rimmele@ae.mpg.de

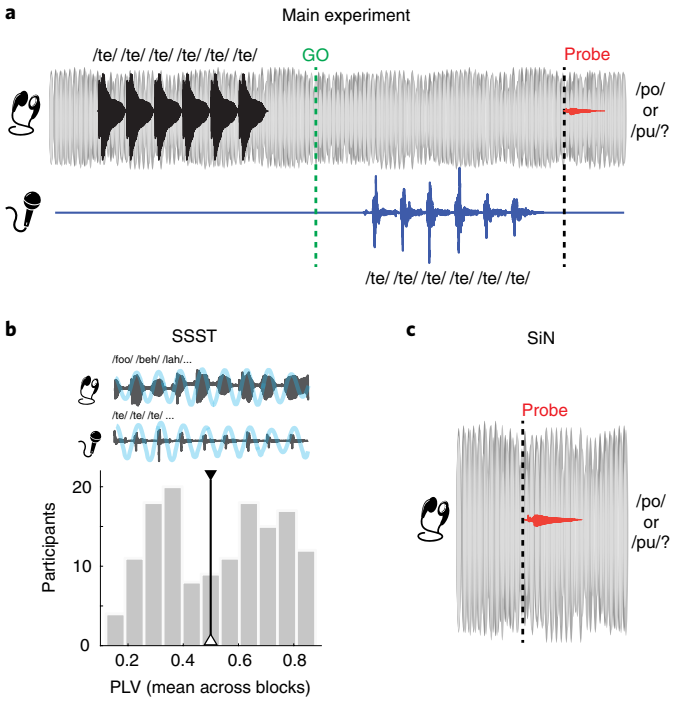


Fig. 1 | Experimental paradigm. a, Schematic description of one trial of the main experiment. First, participants listen to a periodic auditory target syllable sequence (black); next, they wait for a visual 'GO' signal (green dashed line) to vocalize the target sequence with the same rate and number of syllables (blue). Finally, a probe syllable (red; 50% /po/ or /pu/) is presented (onset: black dashed line). At the end of a trial, participants indicate whether they heard /po/ or /pu/. All auditory stimuli are embedded in white noise, and the probe amplitude is presented at participants' individual threshold levels (SiN, 70% correct discrimination). **b**, Before each experiment, following the procedure described by Assaneo and colleagues[33], the spontaneous speech synchronization test (SSS-test) was run. During the SSS-test, participants continuously whisper a syllable (microphone symbol) while concurrently listening to a rhythmic train of syllables (headphone symbol). The histogram of synchronization measurements ($N=143$) displays the phase-locking values between the envelope of the produced and the perceived speech signals (envelopes displayed in blue). Participants on the right (or left) of the black line were categorized as high (or low) synchronizers (highs: $N=75$, lows: $N=68$). **c**, In addition to the SSS-test, before the experiment a SiN test was performed: in each trial, participants listened to background noise (grey), and a probe syllable (red; 50% /po/ or /pu/) was presented (onset: black dashed line) at a random phase of the total trial length. Participants indicated whether they heard /po/ or /pu/. The probe's amplitude was adjusted across trials using Bayesian adaptive estimation in order to get the individual psychometric threshold at 70% correct responses. The headset and microphone icons are used from Stock Unlimited.

Two crucial aspects of the design were carefully controlled. First, as stated before, on the basis of the SSS-test, participants were classified into high or low synchronizer groups that reportedly show significant functional and structural neural differences. The outcome of the present experiment was explored within these groups. Second, the reliability of the observations was assessed across groups and production rates. Two protocols utilizing substantially different production rates were conducted on different samples of participants: one using 'natural' frequencies of speech production and perception (individuals' preferred syllable rate), and another using a fixed 'non-natural' frequency (2 Hz; typical syllable rates lie between 3 and 6 Hz)[21]. Finally, we built a simple neural model consisting of two coupled oscillators (representing speech motor and auditory brain regions) compatible with the behavioural outcome and with the previously reported brain differences between low and high synchronizers.

Results

We conducted two experiments to investigate rhythmic motor-to-auditory entrainment. First, before being assigned to experiment 1 or 2, participants completed the SSS-test (Fig. 1b; see Methods). This behavioural protocol, introduced in previous work[33], splits the population into two groups according to individual differences in the auditory–motor speech synchronization attributes. While concurrently whispering a syllable and listening to a rhythmic auditory stimulus, some participants (highs) spontaneously align their production rate to the perceived one, while others (lows) show no modification in their idiosyncratic rates. Importantly, it has been shown that group affiliation predicts both structural (magnetic resonance imaging (MRI)) and functional (magnetoencephalography (MEG)) features of the brain network supporting speech perception and production[33]. Thus, we decided to assess the outcome of the current studies within each group. Accordingly, as a first step, each participant completed the test and was classified as a high or low synchronizer (Fig. 1b).

Second, individual speech-in-noise (SiN) thresholds were measured. For each participant, the signal amplitude of the probe syllables (/po/ and /pu/), embedded in white noise, was adjusted to 70% correct detection threshold (Fig. 1c).

Third, participants were assigned to main experiment 1 or 2. In both experiments, they completed 150 listen–reproduce–discriminate trials (Fig. 1a). Within each trial, participants first listened to a periodic speech target sequence; next, they waited for a visual 'GO' cue to then reproduce the target sequence; after their speech production ended, a probe syllable (/po/ or /pu/) was presented, and they performed a syllable discrimination task. The probe syllable was presented at the individual SiN threshold level and at a variable time after their speech offset. Importantly, to avoid participants' own auditory feedback, (i) they were instructed to whisper and (ii) white noise was presented during the whole trial. After completing the 150 trials, participants indicated the percentage of the time they perceived their own voice.

Then, we performed the critical analysis: using the produced speech signal we defined a motor oscillation and tested whether the associated phase at which the probe syllable was presented predicted the performance in the discrimination task (Fig. 2 and Methods).

To assess the reliability across samples and speech production rates, each participant completed one of two possible experiments. In experiment 1, motor-to-audio entrainment was assessed at 2 Hz production rate; that is, the target sequence was presented at 2 syllables per second. A frequency of 2 Hz was chosen, as rates close to this value are typically investigated and seem to be optimal for auditory–motor synchronization[35].

Experiment 1 (2 Hz). We first explored the SiN threshold values, the reported own-speech perception and the loudness of the produced speech to ensure that the following results do not derive from differences in these parameters (which are not relevant for the research question). We found no statistically significant difference between low and highs in these parameters, with anecdotal ($1<$ Bayes factor (BF) <3) to moderate ($3<\mathrm{BF}<10$) evidence for the null hypothesis (SiN threshold: Supplementary Fig. 1a, $W=503.0$, $P=0.309$, $r=0.156$, $\mathrm{CI}=[-0.139, 0.426]$, $\mathrm{BF}_{01}=2.31$; $\mathrm{mean}_{\mathrm{highs}}=0.073$, $\mathrm{mean}_{\mathrm{lows}}=0.083$, $n_{\mathrm{highs}}=30$, $n_{\mathrm{lows}}=29$; reported own-speech perception: Supplementary Fig. 1b, $W=445.5$, $P=0.846$, $r=0.024$, $\mathrm{CI}=[-0.266, 0.311]$, $\mathrm{BF}_{01}=3.21$; $\mathrm{mean}_{\mathrm{highs}}=3.50\%$, $\mathrm{mean}_{\mathrm{lows}}=6.6\%$, $n_{\mathrm{highs}}=30$, $n_{\mathrm{lows}}=29$; loudness of the produced speech computed as the root mean square (rms) of the speech recordings: Supplementary

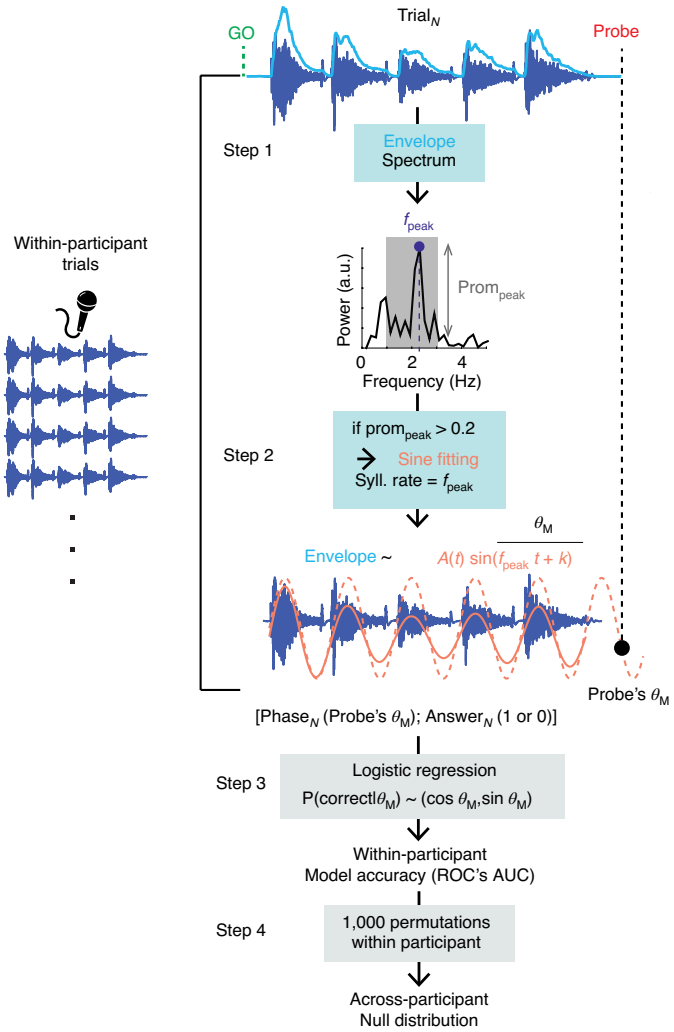


Fig. 2 | Analysis pipeline. For each trial, speech production recordings were processed in steps 1 and 2 to obtain the motor phase at which the probe stimulus occurred. Step 1: the spectrum of the envelope of the produced speech signal between the 'GO' and the probe syllable was computed. The frequency and prominence of the spectrum's peak amplitude (within a ±1 Hz band around the frequency of the target presentation rate) were extracted. Trials with a prominence below 0.2 were rejected. The extracted peak frequency was used as an estimate for the produced syllable rate of this trial and as the frequency for the oscillation that was fitted in the next step. Step 2: a modulated sinusoid was fitted to the envelope (solid orange line), with the phase evolution being continued after voice offset (dashed orange line). The phase corresponding to the probe occurrence was extracted. This process reduced each trial to a two-dimensional vector: a phase and a binary answer (1 for correct and 0 for incorrect). Step 3: for each participant, a logistic regression was fitted on the outcome across trials. The accuracy for each model was computed as the area under the curve (AUC) for the corresponding receiver operating characteristic (ROC) curve. Step 4: to assess the significant presence of information in the adjusted models, a null distribution across participants was computed, by means of 1,000 permutations, shuffling the answers within each participant.

Fig. 1c, $W=457$, $P=0.918$, $r=0.017$, CI=[−0.271, 0.302], $BF_{01}=3.64$; $mean_{highs}=0.041$, $mean_{lows}=0.033$, $n_{highs}=31$, $n_{lows}=29$).

Next we tested whether rhythmic motor production enhances overall detection performance, by comparing the total performance with the SiN 70% correct responses threshold level ($n_{highs}=30$, $n_{lows}=29$). We found evidence for higher overall task performance in the highs ($V=351$, $P_{vs0.7}=0.015$, $r=0.510$, CI=[0.151, 0.750]; $BF_{10}=2.58$; mean 0.73) but not in the lows ($V=212$, $P_{vs0.7}=0.915$, $r=-0.025$, CI=[−0.415, 0.373]; $BF_{01}=5.00$; mean 0.70) (Fig. 3a). There was a trend towards significant differences in the overall performance between groups, albeit with anecdotal evidence for no effect ($3>BF>1/3$) ($W=319$, $P=0.080$, $r=-0.267$, CI=[−0.516, 0.024]; $BF_{10}=0.58$).

Interestingly, we found evidence that, when participants were guided (by the target syllable sequence) to whisper at 2 Hz, both groups produced higher rates than expected (Supplementary Fig. 2, $n_{highs}=30$, $n_{lows}=29$, highs: $V=453$, $P_{vs2Hz}<0.001$, $r=0.948$, CI=[0.887, 0.977], $BF_{10}=2824$, mean 2.09, lows: $V=377$, $P_{vs2Hz}<0.001$, $r=0.733$, CI=[0.477, 0.875], $BF_{10}=555.5$, mean 2.07). No significant difference was found between the group syllable rates with moderate positive evidence ($3<BF<10$) for no difference ($W=408$, $P=0.69$, $r=-0.062$, CI=[−0.345, 0.231], $BF_{01}=3.17$).

Finally, we assessed our main question, namely whether the correct detection of the probe syllable was modulated by the motor phase at which it was presented ($n_{highs}=30$, $n_{lows}=29$). The decoding approach (Fig. 2) provided evidence that the participants' answer was more strongly modulated by the motor phase for the highs than for the lows (Fig. 3b; $W=251$, $P=0.005$, $r=-0.423$, CI=[−0.634, −0.153]; $mean_{highs}=0.590$, $mean_{lows}=0.558$; $BF_{10}=7.25$). Furthermore, we compared the accuracy of the models fitted to the experimental datasets of each group against a null distribution, that is, 1,000 repeated permutation tests within each group. The mean accuracy of the experimental models was above the 95th percentile of the null distribution only for highs, with strong evidence ($10<BF<30$) for the null hypothesis in the lows (Fig. 3b; $95_{prctlHIGH}=0.579$, $mean_{nullHigh}=0.5642$, $BF_{+0}=44.78$; $95_{prctlLOW}=0.578$, $mean_{nullLow}=0.5654$, $BF_{0+}=11.33$). These findings show motor-to-auditory entrainment in high synchronizers, but not in lows. Although the task performance of the highs was modulated by the motor production phase, indicating that the preferred phase of the probe with respect to the motor response was consistent within participants, we found no evidence for an optimal phase across participants (Fig. 3c; omnibus test for non-uniformity, $P=0.67$).

Experiment 2 (individual natural frequency). Experiment 1 revealed that, when speech is rhythmically produced at 2 Hz, it significantly entrains perception only in a subgroup of the population. These findings are compatible with two interpretations. One possibility is that participants with a strong auditory–motor connection, that is, highs, are more flexible with regard to the speech rates they can entrain to. At non-natural speech rates, such as 2 Hz (during fluent speech, typical syllable rates lie between 3 and 6 Hz[21]), the connectivity between motor and auditory regions has been shown to be weaker compared with more typical rates[36]. Thus, it is possible that highs, because of their strong auditory–motor connection, can utilize the motor system better at non-natural rates compared with lows. A different explanation is that the top-down modulation of production to perception is nonexistent, or at least strongly diminished, in a subgroup of the population (lows), regardless of the produced rate. To disentangle these alternatives, we repeated the protocol with a different sample and rate. An additional test was conducted to determine the individual natural speech rate. Participants were primed to speak at 4.5 Hz and subsequently repeatedly whispered the syllable /te/ at their own comfortable rate. The individual speech production signal was recorded and used to compute each participant's natural speech production rate. During the main experiment, this individual rate was used as the presentation rate for the periodic target syllable sequence.

We submitted the data to the same analyses applied in experiment 1. Again, we found no evidence for statistically significant differences between low and high synchronizers in the 'non-relevant'

parameters, with anecdotal evidence for no difference: SiN threshold (Supplementary Fig. 1a; $W=286$, $P=0.298$, $r=0.192$, CI = [−0.159, 0.499], $BF_{01}=2.51$, $mean_{highs}=0.092$, $mean_{lows}=0.097$, $n_{highs}=30$, $n_{lows}=16$), participants' reported auditory feedback (Supplementary Fig. 1b; $W=173$, $P=0.136$, $r=-0.228$, CI = [−0.530, 0.126], $BF_{01}=1.97$, $mean_{highs}=9.9\%$, $mean_{lows}=3.9\%$, $n_{highs}=28$, $n_{lows}=16$) and loudness of the produced speech (Supplementary Fig. 1c; $W=263$, $P=0.486$, $r=0.131$, CI = [−0.224, 0.456], $BF_{01}=2.61$, $means_{highs}=0.055$, $mean_{lows}=0.062$, $n_{highs}=31$, $n_{lows}=15$).

As in the previous experiment, the highs' general performance compared with the 70% threshold level was significant ($n_{highs}=31$, $n_{lows}=16$; $V=360$, $P_{vs0.7}=0.027$, $r=0.452$, CI = [0.083, 0.711]; $BF_{10}=4.86$, mean 0.74). There was no evidence for significant overall detection enhancement by motor production in the lows ($V=36$, $P_{vs0.7}=0.10$, $r=-0.471$, CI = [−0.789, 0.047]; mean 0.65, $BF_{01}=1.1$). This time, highs significantly differed from lows in general performance (Fig. 3a; $W=131.5$, $P=0.009$, $r=-0.470$, CI = [−0.0697, −0.156], $BF_{10}=3.79$). No evidence for significant differences was found when comparing the overall task performance between experiments ($W=1460$, $P=0.642$, $r=0.053$, CI = [−0.168, 0.269], $BF_{01}=4.64$; experiment 1, mean 0.719; experiment 2, mean 0.708; $n_{exp1}=59$, $n_{exp2}=47$).

Interestingly, we found evidence for a significantly slower produced syllable rate during the main experiment for lows compared with highs (Supplementary Fig. 3a; $W=130$, $P=0.012$, $r=-0.458$, CI = [−0.691, −0.140], $BF_{10}=3.18$; $highs_{mean}=4.45$, $mean_{lows}=4.11$; $n_{highs}=30$, $n_{lows}=16$). Such a difference derives from the fact that the participants' natural rate was adopted for the periodic target syllables, with evidence for slower natural rates in lows than highs (Supplementary Fig. 3b; $W=148.5$, $P=0.035$, $r=-0.381$, CI = [−0.638, −0.047], $BF_{10}=1.01$, $mean_{highs}=4.36$ Hz, $mean_{lows}=4.09$ Hz; $n_{highs}=30$, $n_{lows}=16$). This does not imply that lows were primed less successfully by the target sequences during the main experiment, since their production correctly matched the natural rates extracted before the experiment (Supplementary Fig. 3), that is, the rates adopted for the target sequences.

We explored the dependence of the correct detection of the probe syllable on the motor phase at which it was presented ($n_{highs}=30$, $n_{lows}=16$). All the results reported for experiment 1 were replicated: (i) participants' answers were better predicted by the motor phase for the highs than for the lows (Fig. 3b; $W=121.0$, $P=0.005$, $r=-0.496$, CI = [−0.715, −0.187]; $BF_{10}=3.64$; $mean_{highs}=0.595$, $mean_{lows}=0.562$); (ii) the experimental model differed from the null distribution ($mean_{nullHigh}=0.5848$, $mean_{nullLow}=0.5633$) only for highs with strong evidence for H1 (Fig. 3b; $95_{prctlHIGH}=0.579$, $BF_{+0}=204.8$) and moderate evidence for H0 in the lows ($95_{prctlLOW}=0.574$, $BF_{0+}=5.64$); and (iii) it was not possible to define a stable optimal phase across participants (Fig. 3c; omnibus test for non-uniformity of circular data $P=0.92$).

In addition, we run a two-way ANOVA on the total sample of participants ($n=106$; Shapiro–Wilk test: $W(105)=0.993$, $P=0.883$; Levene's test: $F=1.257$, $P=0.293$) to examine the effect of group affiliation (high versus low) and experimental condition (experiment 1 versus 2) on model accuracy. There was no evidence for a significant interaction between the effects of synchrony group and experimental condition on accuracy ($F(1, 102)=2.428$, $P=0.122$, $\eta^2=0.021$). An analysis of simple main effects shows higher model accuracy for highs than for lows ($F(1, 102)=9.784$, $P=0.002$, $\eta^2=0.085$, CI = [−0.100, −0.022]), but no evidence for differences between experimental conditions ($F(1, 102)=1.176$, $P=0.281$, $\eta^2=0.010$, CI = [−0.018, 0.060]).

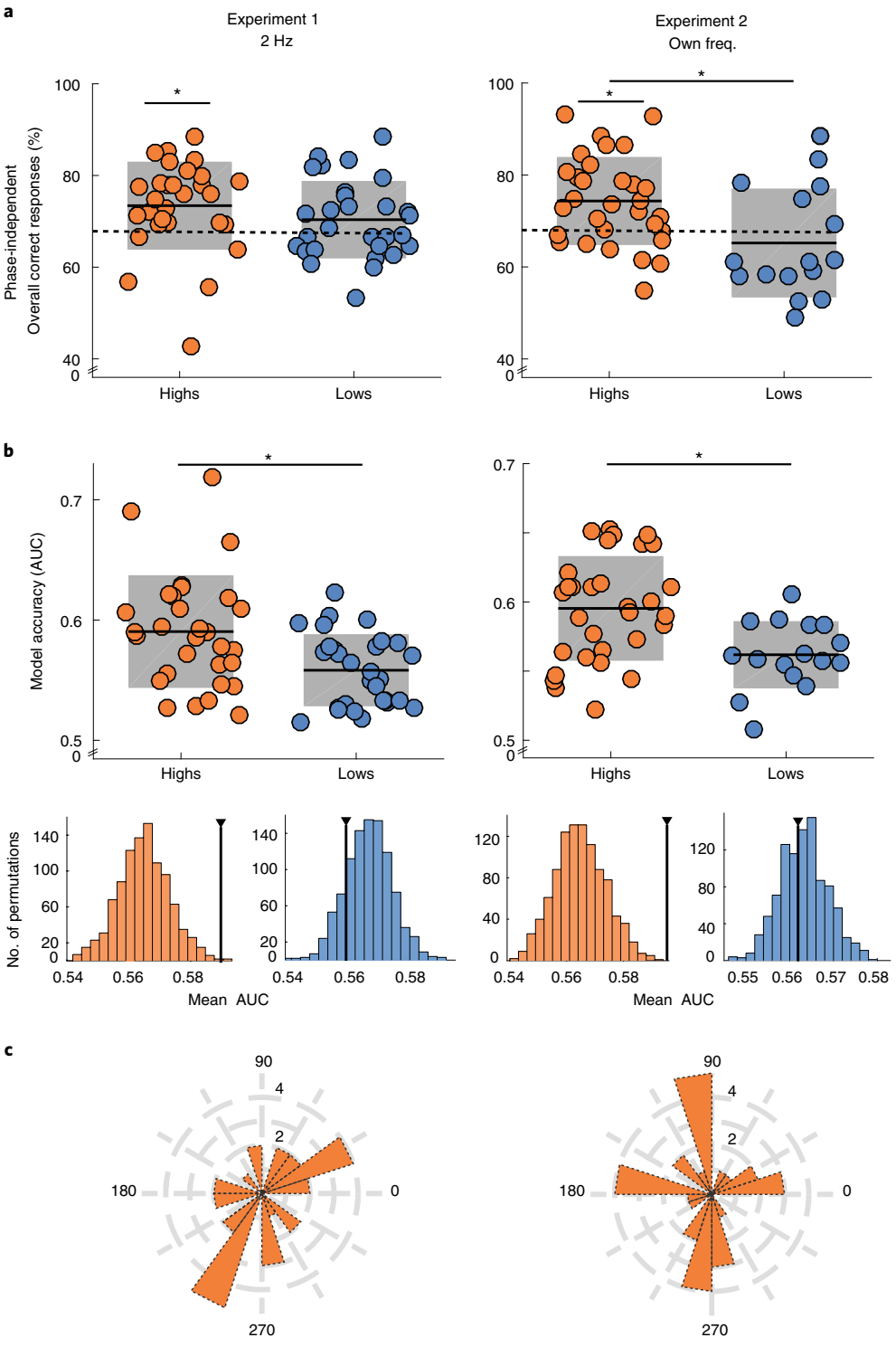
In summary, experiment 2 replicates the outcome of experiment 1. By doing so, it not only shows the reliability of the experimental pattern, but also indicates that the findings can be generalized across production rates. Finally, to further assess the robustness of the results, the predictive power of the fitted models was computed by applying a ten times repeated tenfold cross-validation. This approach—supporting our findings on model accuracy—shows higher predictive power for the models fitted for the highs compared with the lows in both experiments (Supplementary Fig. 4; experiment 1: $W=249$, $P=0.007$, $r=-0.407$, CI = [−0.624, −0.132], $BF_{10}=4.82$, $n_{highs}=30$, $n_{lows}=28$; experiment 2: $W=107$, $P=0.005$, $r=-0.508$, CI = [−0.728, −0.194], $BF_{10}=3.54$, $n_{highs}=29$, $n_{lows}=15$). Additionally, we obtained strong to moderate evidence for predictive power in the lows not exceeding the null distribution (experiment 1: lows $BF_{0+}=11.49$, highs $BF_{+0}=2.6$; experiment 2: lows $BF_{0+}=6.01$, highs $BF_{+0}=80.56$). Given the imbalance in sample size for highs and lows in experiment 2, we additionally run a Monte Carlo simulation[37] to determine the probability of false evidence for H0 given H1. The results of the simulation (10,000 iterations; Cauchy prior $=\sqrt{2}/2$) show that, given the medium effect size we observed (main analysis of experiment 2 for highs: $r=0.656$), the probability of false evidence for H0 with a moderate Bayes factor ($BF_{10}=1/6$) was 0.1%, supporting our interpretation (Supplementary Fig. 5).

Group differences in entrainment are not explained by auditory phase or production periodicity. For several theoretical reasons, it is unlikely that the observed motor-to-auditory entrainment is caused by phase alignment of the auditory probe processing to the auditory target sequence (that is, auditory-to-auditory entrainment). First, a jittered time interval was inserted between the auditory target sequence and the 'GO' cue for starting the motor response, 'breaking' a possible auditory-to-auditory entrainment. Second,

Fig. 3 | Speech production entrains speech perception in high synchronizers. a, General detection performance during the main experiment. In experiment 1 (left column; $n_{highs}=30$, $n_{lows}=29$) the performance of the highs was significantly above 70% ($V=351$, $P_{vs0.7}=0.015$, $r=0.510$, CI = [0.151, 0.750]; $BF_{10}=2.58$; mean 0.73), but not in the lows ($V=212$, $P_{vs0.7}=0.915$, $r=-0.025$, CI = [−0.415, 0.373]; $BF_{01}=5.00$; mean 0.70). There was a trend towards significant differences in the overall performance between groups ($W=319$, $P=0.080$, $r=-0.267$, CI = [−0.516, 0.024]; $BF_{10}=0.58$). In experiment 2 (right column; $n_{highs}=31$, $n_{lows}=16$) highs performed significantly above 70% ($V=360$, $P_{vs0.7}=0.027$, $r=0.452$, CI = [0.083, 0.711]; $BF_{10}=4.86$, mean 0.74) and outperformed lows ($W=131.5$, $P=0.009$, $r=-0.470$, CI = [−0.0697, −0.156], $BF_{10}=3.79$). Dashed line: 70%, threshold value adjusted during the SiN step. **b**, Top row: model accuracy, AUC for a logistic model, assuming that the probability of a correct detection is modulated by the motor phase. The accuracies of the models were significantly better for the highs compared with the lows in both experiments (experiment 1, left column: $W=2.510$, $P=0.005$, $r=-0.423$, CI = [−0.634, −0.153]; $mean_{highs}=0.590$, $mean_{lows}=0.558$; $BF_{10}=7.25$, $n_{highs}=30$ $n_{lows}=29$; experiment 2, right column: $W=121.0$, $P=0.005$, $r=-0.496$, CI = [−0.715, −0.187]; $BF_{10}=3.64$; $mean_{highs}=0.595$, $mean_{lows}=0.562$, $n_{highs}=30$, $n_{lows}=16$). Bottom row: histograms displaying the mean AUC distributions of 1,000 iterations of permuted data within group and experiment. Black lines indicate the experimental mean AUC. The experimental mean AUC was above the 95th percentile computed on the null distribution for the highs but not for the lows (experiment 1, left column: $95_{prctlHIGH}=0.579$, $mean_{nullHigh}=0.5642$, $BF_{+0}=44.78$; $95_{prctlLOW}=0.578$, $mean_{nullLow}=0.5654$ $BF_{0+}=11.33$; experiment 2, right column: mean $AUC_{HIGHs}=0.5848>95_{prctlHIGH}=0.579$, $BF_{+0}=204.8$ and mean $AUC_{LOWs}=0.5633<95_{prctlLOW}=0.574$, $BF_{0+}=5.64$). **c**, In both experiments, there was no evidence that the distribution of optimal phases across high synchronizers significantly differed from a uniform circular distribution (experiment 1, left column: $P=0.56$; experiment 2, right column: $P=0.92$). In all panels: dots represent individual participants; orange and blue represent high and low synchronizers, respectively; $^*P<0.05$; black line indicates the mean value, with the standard deviation shaded.

resonance phenomena have been shown to last up to ~3 cycles and thus would cease due to the elapsed cycles[29]. To empirically rule out this hypothesis, we analysed whether the task performance depends on the phase of the auditory sequence. We repeated the analysis described in Fig. 2 but with the phase of the probe occurrence defined by the auditory target sequences (Supplementary Fig. 6a). As expected, we found no evidence for differences in model accuracies between lows and highs in either of the experiments, with anecdotal evidence for no difference (experiment 1, left column: $W=367$, $P=0.309$, $r=-0.156$, CI=[−0.426, 0.139]; $BF_{01}=2.111$, $n_{highs}=30$, $n_{lows}=29$; experiment 2, right column: $W=237$, $P=0.955$, $r=-0.012$, CI=[−0.351, 0.329], $BF_{01}=3.18$, $n_{highs}=30$, $n_{lows}=16$). The non-permuted mean AUC remained below the 95th percentile computed on the null distribution ($mean_{nullExp1}=0.5629$, $mean_{nullExp2}=0.5647$) for both experiments (experiment 1, left column: mean AUC=0.569>95_{prctl}=0.573, $BF_{0+}=1.77$; experiment 2: mean AUC=0.571>95_{prctl}=0.576, $BF_{0+}=2.6$; Supplementary Fig. 6b,c).

Still, the findings of group differences in entrainment are compatible with two different hypotheses: (i) the one advanced in this work: highs and lows significantly differ in their motor-to-auditory coupling; and (ii) an alternative one: the ability to produce periodic



motor patterns is reduced in lows. Decreases in periodicity of the produced sequences would decrease its ability to entrain the auditory system, and furthermore increase the noise in the motor phase measure, resulting in a non-significant outcome. To further explore the origin of the group differences in entrainment, we compared the degree of periodicity of the motor production between groups. We found no significant difference in periodicity, with weak evidence for no difference (experiment 1, $W=510$, $P=0.680$, $r=0.063$, CI = [−0.223, 0.339], $\text{mean}_{\text{highs}}=0.047$, $\text{mean}_{\text{lows}}=0.049$, $\text{BF}_{01}=3.38$, $n_{\text{highs}}=32$, $n_{\text{lows}}=30$; left panel: experiment 2, $W=216$, $P=0.392$, $r=-0.156$, CI = [−0.468, 0.190], $\text{BF}_{01}=2.57$, $n_{\text{highs}}=32$, $n_{\text{lows}}=16$), ruling out the second hypothesis (Supplementary Fig. 7).

Auditory-to-visual entrainment: reaction time analysis. The time interval between the target syllables' offset and the visual 'GO' cue was randomly assigned for each trial to one of the following fractions of the corresponding target period: 1/2, 1, 3/2, 2, 5/2. In other words, it was presented in or out of phase with respect to the auditory stimulus (Fig. 4a). Here, we explore whether the reaction time (RT; that is, the interval between the 'GO' cue and speech onset, Fig. 4a) was modulated by the phase of the 'GO' cue presentation and whether this modulation differed between groups. The aim is to assess whether the difference between highs and lows in motor-to-auditory entrainment extended to other sensory modalities, that is, in this case auditory to visual. Importantly, this is an exploratory analysis; the current experimental design was not optimized to explore the auditory-to-visual entrainment. The analysis serves as a first step to further characterize the sensitivity and specificity of the reported phase effects.

First we compared the mean RT across all trials between groups. Overall RTs were faster in highs compared with lows in experiment 2 ($W=480$, $P=0.005$, $r=0.455$, CI = [0.165, 0.672]; $\text{BF}_{10}=5.54$; $\text{mean}_{\text{highs}}=0.380$, $\text{mean}_{\text{lows}}=0.483$, $n_{\text{highs}}=33$, $n_{\text{lows}}=20$). In experiment 1, there was no significant difference between groups, with anecdotal evidence for no difference (Fig. 4b; $W=2{,}382$, $P=0.209$, $r=0.128$, CI = [−0.071, 0.317]; $\text{BF}_{01}=2.25$; $\text{mean}_{\text{highs}}=0.447$, $\text{mean}_{\text{lows}}=0.478$, $n_{\text{highs}}=33$, $n_{\text{lows}}=32$). The fact that group RT differs only in experiment 2 suggests that this distinction derives from the variability in the target syllable rate, which was significantly slower for the lows compared with the highs only in the second protocol (Supplementary Fig. 3b). Next, we examined the normalized RTs as a function of the time between the target syllable offset and the 'GO' cue. For both experiments, the data displayed a clear anticipation effect: faster RTs for longer delays, as typically reported in the literature[38–40] (Fig. 4b).

We adjusted an exponential function to each individual dataset and submitted the residuals to a decoding approach: exploring whether the residuals allowed to infer the phase (in or out of phase with respect to the auditory rhythm) of the 'GO' cue (Fig. 4c and Methods). This allows to investigate whether the RTs exhibited, in addition to the overall anticipatory effect, auditory-to-visual entrainment. For both experiments there was no significant difference between groups, with moderate evidence for no difference, on how well a logistic regression differentiates the in/out-of-phase status of the visual 'GO' cue (Fig. 4d; experiment 1: $W=537$, $P=0.911$, $r=0.017$, CI = [−0.260, 0.291]; $\text{BF}_{01}=3.76$; $\text{mean}_{\text{highs}}=0.570$, $\text{mean}_{\text{lows}}=0.568$; experiment 2: $W=329$, $P=0.993$, $r=-0.003$, CI = [−0.316, 0.310]; $\text{BF}_{01}=3.45$; $\text{mean}_{\text{highs}}=0.551$, $\text{mean}_{\text{lows}}=0.548$). Thus, the whole population was compared against a null distribution (1,000 permutation iterations within experiment). Crucially, in both experiments, the model performed better in inferring the phase of the visual 'GO' cue compared with the null distribution (Fig. 4e; experiment 1: mean $\text{AUC}_{\text{all}}=0.569>95_{\text{prctl}}=0.547$, $\text{BF}_{+0}=210.6$; experiment 2: mean $\text{AUC}_{\text{all}}=0.550>95_{\text{prctl}}=0.5461$, $\text{BF}_{+0}=1.149$), suggesting that auditory-to-visual entrainment occurred in all participants and experiments.

A neural model grounded on previous research explains the behavioural observations. In this section, we used mathematical modelling and numerical simulations to further validate the hypothesis that rhythmic speech production can modulate perception.

We represented auditory and motor cortices and their interaction as a minimal system of two coupled phase oscillators with noisy, intrinsic preferred frequencies and with delayed mean-field coupling[41]. Both speech motor[36,42] and auditory regions[43] have been previously represented by a neural oscillator, but in contrast to previous approaches, we add a bidirectional interaction between areas with variable strength across individuals.

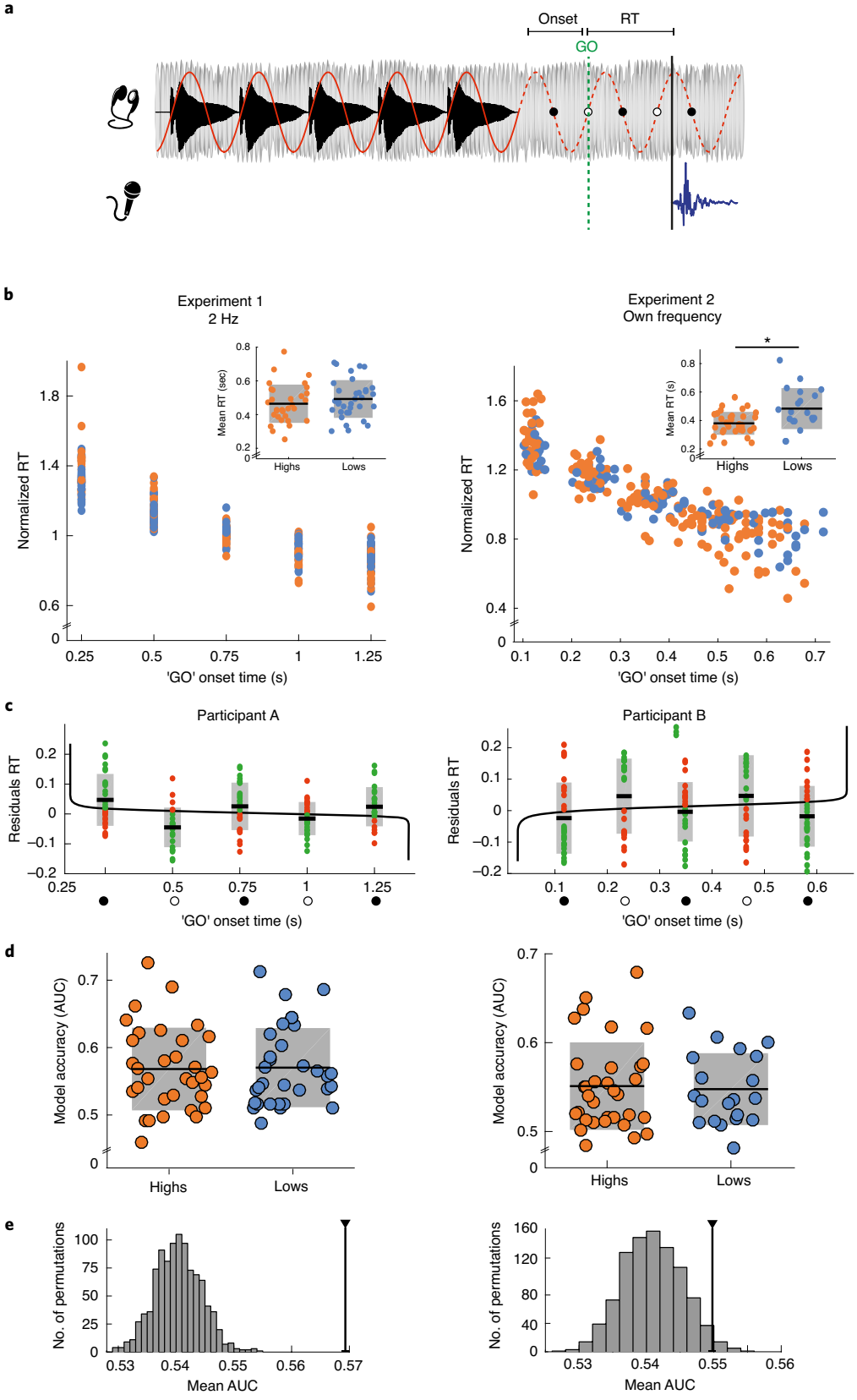
Previous work[33] shows that highs have more volume in the left arcuate fasciculus—the main structural connection between auditory and motor areas—than lows. On the basis of this result, we scale the bidirectional coupling strength between oscillators with the synchrony measurement obtained with the SSS-test. In addition, we link the motor cortex rhythmic activity to the temporal structure of the speech envelope, which represents our experimental measurement. We built on two previous studies showing that: (i) motor cortex activity correlates with the produced syllable rate[44] and (ii) the time delay between mouth movement and speech onset varies by about 100 to 200 ms[45]. On the basis of these observations, we model the recorded syllabic rate as the rhythmic motor output with a variable phase lag (Fig. 5a and Methods).

With this model we run numerical simulations of both experiments. We computed the synchronization between the activity of auditory and motor cortices while participants speak at about 2 (exp. 1) and 4.5 Hz (exp. 2) with no auditory feedback, by calculating the phase-locking value (PLV) between oscillators. The auditory intrinsic frequency was obtained for each participant and for each simulation from a Gaussian distribution centred at 4 Hz, following previous work[43]. For the motor frequency, the centre of the Gaussian

Fig. 4 | Speech perception entrained visual perception in high and low synchronizers. a, Schematic representation of the variables used for the analyses. Top: the auditory stimulus. Bottom: the beginning of the produced utterance. Solid red line: envelope of the acoustic signal filtered (±1 Hz) around the syllable rate. Dashed red line: continuation of the periodic envelope. Green, 'GO' cue; grey line: vocalization onset. Dots represent the possible positions for the 'GO' cue (white in phase and black out of phase). RT, reaction time. **b**, In both groups and experiments (left and right columns, respectively), the normalized RTs are shown as a function of the time delay between the target offset and the 'GO' cue show an exponential decrease. Inset: mean RT across all trials. Highs show faster RT than lows in experiment 2 (right; $W=480$, $P=0.006$, $r=0.455$, CI = [0.165, 0.672]; $\text{BF}_{10}=5.54$; highs mean: 0.380, lows mean: 0.483; $n_{\text{highs}}=33$, $n_{\text{lows}}=20$), but not in experiment 1 (left; $W=2{,}382$, $P=0.209$, $r=0.128$, CI = [−0.071, 0.317]; $\text{BF}_{01}=2.25$; $\text{mean}_{\text{highs}}=0.447$, $\text{mean}_{\text{lows}}=0.478$, $n_{\text{highs}}=33$, $n_{\text{lows}}=32$). **c**, For experiment 1 (left column) and experiment 2 (right column), an exponential function was fitted for each participant, and the residual RTs are plotted (here exemplarily for one participant). A logistic regression was adjusted on the residual RT to predict whether the 'GO' cue appeared IN or OUT of phase (black line). Red and green, respecitvely, show incorrectly and correctly classified trials. **d**, In both experiments, the model accuracy did not distinguish between highs and lows (experiment 1: $W=537$, $P=0.911$, $r=0.017$, CI = [−0.260, 0.291]; $\text{BF}_{01}=3.76$; $\text{mean}_{\text{highs}}=0.570$, $\text{mean}_{\text{lows}}=0.568$; experiment 2: $W=329$, $P=0.993$, $r=-0.003$, CI = [−0.316, 0.310]; $\text{BF}_{01}=3.45$; $\text{mean}_{\text{highs}}=0.551$, $\text{mean}_{\text{lows}}=0.548$). **e**, Histograms displaying the values for the AUC of 1,000 permutations of the data shuffling the IN and OUT labels within participants averaged across the whole cohort of each experiment. Black lines indicate the experimental mean AUC. The experimental mean AUC was above the 95th percentile computed on the null distribution for both protocols (experiment 1: mean $\text{AUC}_{\text{all}}=0.569>95_{\text{prctl}}=0.547$, $\text{BF}_{+0}=210.6$; experiment 2: mean $\text{AUC}_{\text{all}}=0.550>95_{\text{prctl}}=0.5461$, $\text{BF}_{+0}=1.149$).

varied with the experimental condition: 2 Hz for experiment 1 and 4.5 Hz for experiment 2.

Numerical simulations provide evidence for an enhancement of auditory–motor synchronization for highs compared with lows (Fig. 5b; experiment 1, $W=158$, $P<0.001$, $r=-0.637$, CI = [−0.782, −0.426], $BF_{10}=62.7$, $n_{highs}=30$, $n_{lows}=29$; experiment 2, $W=116$, $P=0.002$, $r=-0.532$, CI = [−0.738, −0.237], $BF_{10}=7.57$, $n_{highs}=31$, $n_{lows}=16$). To evaluate the statistical significance of the simulated results, we performed 1,000 runs for each experiment. For each run, we computed the P value of the PLV comparison between highs and lows.



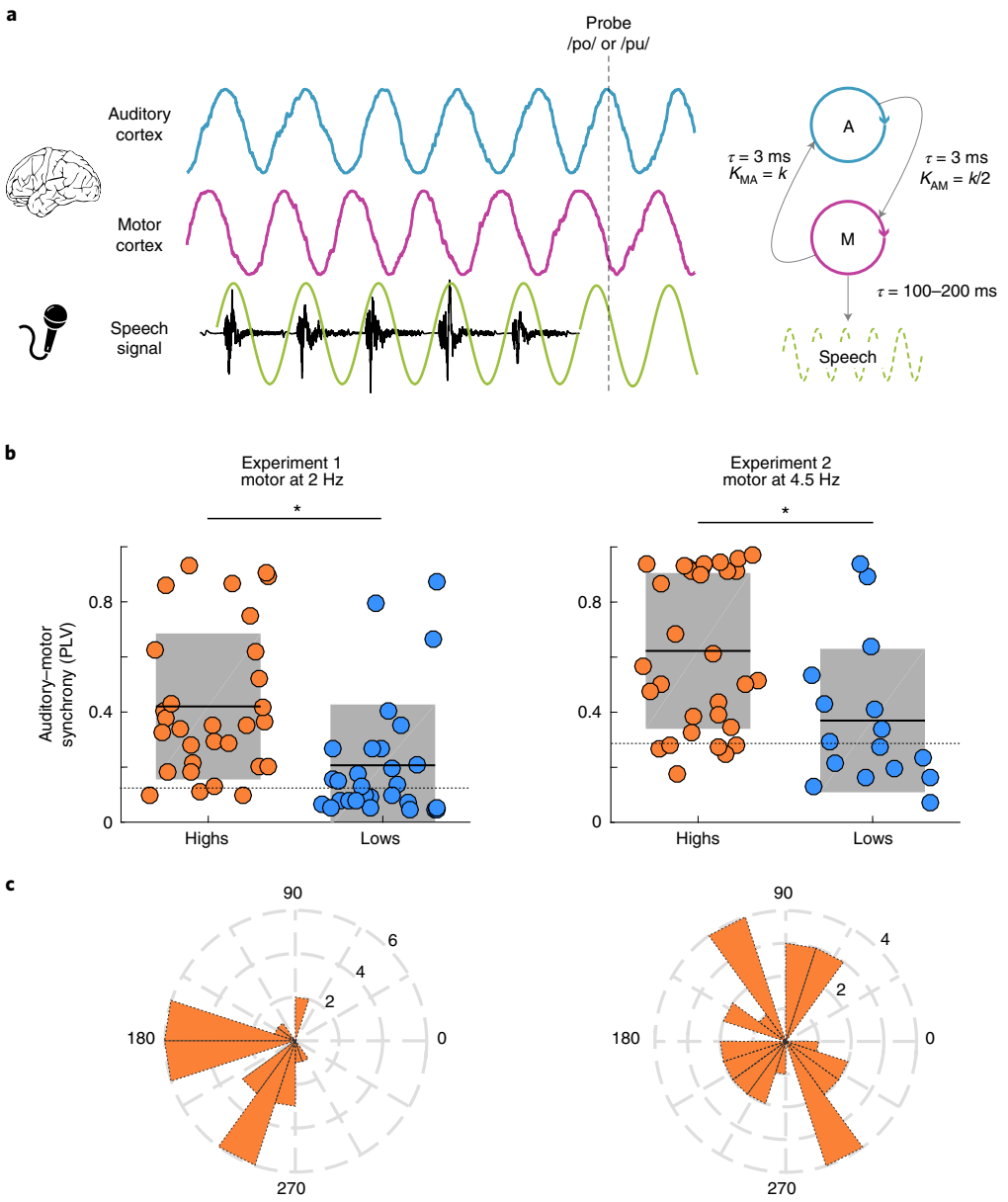


Fig. 5 | Neural model compatible with the behavioural observations. a, Schematic representation of the model. Speech motor and auditory cortices activity (blue and magenta traces, respectively) are modelled as a set of coupled phase oscillators. The oscillatory dynamic of the produced speech is modelled as the motor cortex activity with a fixed phase lag representing the time delay observed between muscle activity and the onset of the produced speech[45]. The headset and microphone icons are used from Stock Unlimited. The brain icon is used from Pixabay. The auditory–motor interaction schematic is adapted from ref. [36]. **b**, Synchronization between the auditory and motor cortices activities obtained by simulating the different experimental conditions with the numerical model. Dots represent individual participant's data, obtained by scaling the strength of the interaction between the oscillators according to the participant's degree of speech auditory–motor synchrony assessed with the SSS-test. The obtained PLVs were significantly different between high and low synchronizers (experiment 1, left column: $W = 158$, $P < 0.001$, $r = -0.637$, CI = [−0.782, −0.426], $BF_{10} = 62.7$, $n_{highs} = 30$, $n_{lows} = 29$; experiment 2, right column: $W = 116$, $P = 0.002$, $r = -0.532$, CI = [−0.738, −0.237], $BF_{10} = 7.57$, $n_{highs} = 31$, $n_{lows} = 16$). The dashed line represents the baseline phase-locking value for each experimental condition, obtained by averaging the PLVs of 100 iterations for each experimental condition with no interaction between the oscillators ($K_{AM} = K_{MA} = 0$). Orange and blue, high and low synchronizers, respectively; $^*P < 0.05$; black line, the mean value with the standard deviation shaded. **c**, Distribution of the mean phase lag between the auditory cortex activity and the produced speech envelope (blue and green traces in **a**, respectively). For both experimental conditions, the distribution of phase lags did not significantly differ from a uniform circular distribution (experiment 1, left column: $P = 0.15$; experiment 2, right column: $P = 0.96$).

The PLV difference between groups was significant for 96.6% of the iterations in experiment 1, and for 95.4% in experiment 2. Enhanced auditory–motor synchronization in the highs implies a more stable phase lag between the produced syllables sequences (Fig. 5a, green trace) and auditory activity (Fig. 5a, blue trace) across trials. The modelling results are compatible with our behavioural findings, where we observed better syllable discrimination performance for a given recorded motor phase. In other words, if motor and auditory activity are time locked, the phase of high auditory cortex excitability aligns to the same 'speech' phase across trials within participants. Assuming individual variability not only in the coupling strength, but also in the intrinsic frequency of auditory cortex and the time delay

between muscle activity and speech onset, results in a non-uniform distribution of phase lags across participants. In other words, each high synchronizer individually shows constant speech production to auditory phase lag. However, the value of the constant lags varies across individuals (Fig. 5c). This outcome can explain why we found no evidence for an optimal phase across highs. Crucially, in case an external rhythmic auditory stimulus is presented (which was not the case in our study), the model predicts a tighter alignment of audio–motor phase delays across participants (Supplementary Fig. 8).

Discussion

The current study shows that rhythmic speech production entrains perception, which suggests the involvement of oscillatory processes in the auditory–motor interaction. A critical finding is that motor-to-auditory entrainment was observed only in part of the population, suggesting that individuals differ in their ability to exploit rhythmic motor activity to support speech perception. Interestingly, it was not possible to define a stable optimal phase across participants. The individual differences highlight the benefit of our analysis approach, which in contrast to other methods is flexible with respect to such individual variance. The findings provide mechanistic insight into the auditory–motor integration of speech: in accordance with oscillatory regimes of motor and auditory cortex, we show that a simple neural model in which both areas are represented by a set of interacting phase oscillators can explain the observed behavioural pattern.

What might account for the observed variability in entrainment effects? We used a recently validated behavioural test, the SSS-test, to divide the population according to individual differences in their auditory-to-motor synchronization. Note that the replicability of the bimodal distribution of individuals' spontaneous speech synchronization was reported within the original study[33], and is further supported by our findings, which replicate the distribution in a sample of native German speakers (Fig. 1a). Crucially, perceptual entrainment to the speech production rhythm was only observed for participants identified as high speech synchronizers. Thus, our findings provide direct behavioural evidence for a bidirectional effect: participants who show a strong spontaneous synchronization of their speech production to an external auditory speech input (that is, auditory to motor), also show significant top-down effects from the motor system to enhance perception (that is, motor to auditory). Note that the findings were not underpinned by speech-in-noise threshold differences, nor by residual auditory perception: the overall percentage of participants perceiving their own speech was low, and there was no evidence for group differences between high and low synchronizers, nor in the loudness of their own speech production. Furthermore, findings cannot be explained by auditory-to-auditory entrainment, nor by group differences in motor production periodicity.

Previous research showed increased structural connectivity between temporal and frontal regions in the brains of high synchronizers[33]. On the basis of this finding, we hypothesize that the same structural pathways support motor-to-auditory entrainment. To further test this hypothesis we run numerical simulations of our experimental design using a simple neural model in which the strength of the interaction between auditory and motor cortices scales with the individual speech-to-speech synchronization measurement (which, as stated above, is predictive of the left arcuate fasciculus volume[33]). We found that the numerical simulations obtained using the model align well with the observed behavioural pattern.

A possible alternative explanation is that, even though participants could not hear their own speech, they might have imagined it, resulting in mere auditory entrainment[46]. However, this would imply that high synchronizers have a stronger imagery effect than lows, which—in contrast to our proposed model—does not derive from the brain structural difference between groups.

Regarding the phase lag of perception–production entrainment, previous findings are controversial. On the one hand, auditory processing of self-generated (and thus fully predictable) sound during speech production is typically suppressed compared with external input processing[47], possibly reflecting anti-phase entrainment of auditory neuronal oscillations to the low-excitability phase of the motor cortex activity[48,49]. On the other hand, for external input predicted by a motor action, neuronal activity is typically enhanced compared with non-predicted input[50]. Most crucially, knowledge of the behavioural consequences of perceptual entrainment is scarce. In our study, while the optimal phase for auditory perception was stable within participants, it varied across participants (Fig. 3c). Variations in optimal phase across participants have been reported for behavioural, neurophysiological and transcranial alternating current stimulation (tACS) studies[51–55] (for methodological aspects, see ref. [56]). Here, numerical simulations show that a bidirectional excitatory connection between auditory and motor cortices can explain the observed phase lag pattern when individual differences are incorporated into the model. Furthermore, the model sheds some light on an apparent contradiction: how can the entrainment be beneficial if it is not possible to define an optimal phase lag across participants? The confusion here is that in our experimental design there is no objective optimal phase, giving that the external stimulus presentation does not follow any rhythmic pattern (that is, the probe is presented at a random phase at each trial). In the case of having a rhythmic external stimulus, the model predicts a tighter alignment for the auditory–motor phase lag across individuals (Supplementary Fig. 8).

The finding of substantial individual differences in the motor-to-auditory entrainment suggests that rhythmic top-down predictions from the motor system are not equally used for speech perception across the population. A natural follow up question is whether the individual differences in entrainment extend to other sensory modalities or are restricted to the speech motor-to-auditory interaction. Although our design was not optimized to explore systematically the entrainment across different sensory modalities, an analysis of auditory-to-visual entrainment revealed evidence for effects in all participants. This suggests that individual differences do not extend to auditory-to-visual sensory entrainment[35,57].

The experiments we describe provided evidence for a critical role of temporal predictions arising from the speech production motor system during speech perception, on the basis of oscillatory processes. Importantly, the reported effect was measurable only in parts of the population. The tight coupling between the auditory and motor systems is further supported by the specificity of the effects for auditory perception. The findings have significant implications for models of speech processing, inviting among other issues models that incorporate individual differences in a more principled way.

Methods

Participants. The experimental procedures were ethically approved by the Ethics Council of the Max Planck Society (no. 2017_12). Before the study, all participants gave written informed consent. Monetary compensation was provided. Data collection and analysis were performed blind to whether participants were high or low synchronizers. An initial cohort of 143 participants was included in the SSS-test analysis (highs: $N=75$; lows: $N=68$). Because of no consistency between the synchrony measurement (PLV) during the first and second block of the SSS-test, several participants were excluded before the analysis ($N=6$). After the SSS-test, each participant was assigned to complete one of the experiments. No statistical methods were used to pre-determine sample sizes, but our sample sizes are similar to those reported in a previous publication using similar protocols[33].

In experiment 1, the data of 59 (highs: $N=30$; mean age: 24.38 years, s.d.: 3.42 years; female: 44) participants were included in the analysis. Several participants were excluded before the analysis for different reasons: difficulties in following the instructions ($N=3$), data loss ($N=1$) and fewer than 100 trials surviving the cleaning process in the main experiment ($N=8$; highs: $N=2$). For all analyses, participants with outlier data were removed (exceeding 2 standard deviations).

In experiment 2, the data of 46 (highs, $N=30$; mean age: 25.18 years, s.d.: 4.19 years; female: 32) participants were included in the results. Several participants

were excluded from the experiment before the analysis, because their natural syllable rate lay outside the range of 3.5–5.5 Hz ($N=17$; highs, $N=7$), or because, after the preprocessing, the number of trials in the main experiment was below 100 ($N=7$; highs: $N=2$).

Stimulus processing. The syllables /po/, /pu/ and /te/, slowly spoken by a native German speaker, were recorded in a sound-attenuated booth with an external audio card (Brand; 44,100 Hz/16 bits) as digital input and a customized MATLAB script using Psychophysics Toolbox extensions[58] and PRAAT software (http://www.praatvocaltoolkit.com/)[59]. The recorded syllables were high-pass filtered at 60 Hz to remove background noise. Syllables were normalized in peak-amplitude and pitch contour (at 220 Hz). For each syllable, the duration was adjusted according to the syllable presentation rate (exp. 1: 0.5 s to generate a 2 Hz rate; exp. 2: individually adjusted to match the period corresponding to the target frequency).

Experimental procedures. All experimental parts of the study were conducted in a sound-attenuated booth. Participants were seated at a personal computer with a microphone (MX418 gooseneck microphone, Shure) in front of their mouth (cotton sticks were used to keep the distance constant at 3 cm). The sound was presented through ear-plugs (E-A-RTONE Gold 3 A insert earphones, Ulrich Keller Medizin-Technik, Weinheim, Germany). A keyboard was used for the behavioural response. A volume control knob allowed for loudness adjustment. In all cases, where a loudness adjustment was performed, participants were instructed to increase the volume until their own whispering became inaudible, up to a maximum loudness level that was still comfortable. The procedure allows to investigate the effects of speech–motor processing on speech perception, while the auditory perception of one's own speech is suppressed.

Spontaneous speech synchronization test (SSS-test). In the SSS-test, participants performed the following steps (for a detailed description of the procedure see ref. [33]): first, participants adjusted the loudness of a 'babble noise' (syllables played backwards), while listening to the babble noise and at the same time continuously whispering the syllable /te/. Second, participants performed two training trials. They were instructed to listen to a continuous presentation of the syllable /te/ (presented at 4.5 syllables per second) for 10 s, and afterwards repeat the syllable by continuously whispering at the same rate (10 s). Third, during the test trials, participants listened to a continuous sequence of random syllables (for example, 'foo beh lah etc.') at a rate of 4.5 syllables per second for 1 min. They were instructed to pay attention to the syllable sequence in order to perform a syllable discrimination task thereafter and at the same time—in order to 'increase the task difficulty'—continuously whisper the syllable /te/. During all trials, a fixation cross was presented. After each trial, syllables were presented and participants indicated whether a syllable was part of the sequence. The second and third step were performed twice (run 1 and run 2).

Speech-in-noise threshold test. To individually adjust the signal-to-noise ratio in the main experiment, a Bayesian adaptive estimation of the individual psychometric threshold of the speech-in-noise perception was conducted (using the Palamedes toolbox[60]). First, participants adjusted the loudness of a white noise stimulus while they were continuously whispering the syllable /te/. (Note that the loudness level was kept constant for all following parts of the study.) Second, the speech-in-noise perception threshold was estimated. The signal-to-noise ratio between the noise and a speech probe amplitude was adapted trial-wise (prior SNR threshold range in dB: −32:1:−10; prior slope: −1:0.0263:20; prior chance rate: 0.5; prior lapse range: 0:0.01:0.06; signal-to-noise ratio value range in dB: −32:0.02:−10). In the beginning of each trial, a fixation cross was presented, simultaneously with white noise, and after a random time interval between 0.25 and 1.25 s a target syllable was presented (50% /po/ or /pu/) embedded in the noise. At the end of the trial, participants performed a syllable discrimination task, indicating whether they heard the syllable /po/ or /pu/ (button 1 or 2). Feedback ('correct', 'wrong') was given throughout the threshold measure. In the beginning participants performed several training trials (maximum ten trials, terminated after four correct trials) with a fixed SNR (0.8). After the training, a first run started with an easy SNR value. The second run started with a more difficult SNR value. In the second run, the slope estimate of the first run was used as prior in order to increase the precision of the slope estimation. Each run had a minimum of 50 trials and stopped after 12 reversals, that is, changes in correct/error responses. The 70% threshold was selected on the basis of the psychometric function and averaged across runs.

Main experiment. In both experiments, participants were instructed to perform the following sequential steps: (i) listen to a rhythmic target containing a sequence of syllables (for example, /te/ /te/ /te/ /te/ /te/), (ii) whisper the target sequence at the same rate and with the same amount of syllables as soon as a visual go signal (green 'GO') was presented, (iii) listen to a probe syllable presented thereafter (/po/ or /pu/, 50% each) and (iv) at the end of the trial, discriminate which target syllable was heard, /po/ or /pu/. In the beginning of each trial a fixation cross was presented, simultaneously with a white noise stimulus that was present throughout the whole trial. All stimuli were embedded in the noise, and importantly, the amplitude of the probe syllable was adjusted according to the previously estimated 70% threshold (see the "Speech-in-noise threshold test" section).

The target syllable sequence consisted of five or six /te/ syllables (randomly selected, 50% each) presented at a fixed rate, f_{target}. The 'GO' signal was presented after a randomized interstimulus interval corresponding to different phases of the target oscillation: 0.5, 1, 1.5, 2, 2.5 times the period ($1/f_{target}$). The probe syllable was presented (at the individual SNR) randomly within different intervals after the 'GO'. The time intervals were 5.25–7.75 times the target period, for targets with five /te/ syllables, or 6.25–8.75 times the target period for targets with six /te/ syllables. The time intervals were selected to maximize the probability of the probe occurring within three cycles after the speech production offset.

In experiment 1, f_{target} was set to 2 Hz and kept constant across participants. In experiment 2, it was individually adjusted according to the participant's natural rate.

Thus, in experiment 2 the individual natural rate was estimated in a procedure before the SiN threshold test (the loudness level was adjusted before this measure and kept constant throughout the study). During this step, participants were first presented with an example audio of a constant speech production rate (/te/ syllables presented at 4.5 Hz for 10 s) and second asked to whisper continuously and with a consistent rate the syllable /te/ at a comfortable rate (1 min recording duration). The procedure was repeated six times (the first run was considered a training run and removed from the analysis). The mean speech production rate across trials was computed by, first, computing the speech envelope. Second, the spectrum was computed for each trial on the basis of the Fourier-transformed speech envelopes. Finally, the natural rate was computed as the frequency with maximal power of the mean spectrum across trials. Only participants with a mean rate within the range of 3.5–5.5 Hz were further tested. We focused on the theta range, as it has been previously reported to be optimal for the coupling of the audio–motor cortex during speech perception[36].

Both experiments had 150 trials with a short break after half of the experiment. Before the experiment, participants performed a short training (three trials) where the target was presented at a high SNR (0.8) and feedback was given ('correct', 'wrong'). No feedback was given during the experiment.

At the end of each experiment, participants indicated whether they heard their own speech, and given a positive answer the percentage of time this was the case.

Analyses. For every analysis, the speech envelope was computed as the absolute value of the Hilbert transform of the acoustic signal. For the cochlear envelope, the spectrograms of the auditory stimuli were computed using the Neural System Laboratory (NSL) Auditory Model MATLAB toolbox[61,62] and auditory channels between 180 and 7,246 Hz were added. All envelopes were resampled at 100 Hz and detrended, and their phases were extracted by mean of the Hilbert transform.

Phase-locking value (SSS-test). A measurement of the spontaneous speech-to-speech synchronization was obtained for each participant following ref. [33]. For each run of the SSS-test, the PLV was estimated according to the following formula:

$$PLV = \frac{1}{T}\left|\sum_{t=1}^{T} e^{i(\theta_1(t)-\theta_2(t))}\right|$$

where t is the discretized time, T is the total time points within a 5 s window, θ_1 is the produced speech envelope and θ_2 is the cochlear envelope of the auditory stimulus, with both envelopes filtered between 3.5 and 5.5 Hz. The overlap between windows was set to 2 s, and the PLVs of all windows were averaged within each run. The mean PLV across the first and the second run was defined as the participant's speech synchrony.

Participants were assigned to the low- or high-synchrony group by defining a threshold by applying a k-means algorithm with two clusters to the total distribution of mean PLVs. In a previous work, the bimodal structure of the distribution was shown in a larger sample[33]. On the basis of this result, we adopted a k-means algorithm with two clusters to standardize the group classification.

Main experiment analysis. Automatic preprocessing. For each trial, the envelope was computed for the produced speech signal between the 'GO' cue and the presentation of the probe syllable, and the fast Fourier transform was used to estimate the envelope's spectrum. From the spectrum the produced syllable rate, f_{prod}, was estimated as the peak within a range of $f_{target} \pm 1$ Hz. Trials were removed if: (i) the peak prominence was below 0.2 (that is, participant the did not manage to rhythmically whisper at a certain frequency), or (ii) the participants produced an incorrect amount of syllables ($N_{targetSyll.} \neq N_{prod.Syll}$). Participants with fewer than 100 good trials were removed from the experimental analysis.

Next, for each trial the following function was fitted to the produced speech envelope: $A(t)\sin(f_{prod}t + \theta)$. The corresponding motor phase of the probe presentation is extracted from the continuation of the sine wave fitting (Fig. 2, steps 1 and 2).

Decoding analysis. For each participant's dataset we adjusted a logistic regression where the correct response probability (1 for correct, 0 for incorrect) is modulated by the phase at which the probe was presented, $P(\text{correct}|\theta_M) \propto [\sin\theta_M, \cos\theta_M]$. The accuracy of each logistic regression was computed as the area under the

curve (AUC) of the receiver operating curve computed for the logistic regression outcome, given the true correct or incorrect response behaviour of participants[63,64].

To test the significance of the model's accuracies, we ran repeated permutation tests in which the trial assignments of the probe phase were kept and the behavioural responses (1 correct, 0 incorrect) were randomly shuffled across trials within participant. The testing consisted of 1,000 permutations. The mean accuracy (mean AUC) was computed for each permutation, resulting in a distribution of 1,000 values. Finally, the model's accuracy was deemed significant if it exceeded the 95th percentile of the distribution of the mean AUCs based on shuffled answers.

Statistical analyses. All comparisons between groups were assessed using non-parametric Mann–Whitney *U* tests (*W*) for independent-sample tests and Wilcoxon signed-rank tests (*V*) for one-sample tests. All *P* values reported in this work are two-sided, effect sizes are reported using the rank-biserial correlation (*r*) and CI stands for the 95% confidence interval.

For the exploration of the preferred phase distribution, an omnibus or Hodges–Ajne test for non-uniformity of circular data was computed with H0: the population is uniformly distributed around the circle and HA: the population is not distributed uniformly around the circle.

Bayes factors BF_{01}, which reflect how likely data are to arise from the null model (that is, the probability of the data given H0 relative to H1), and BF_{10}, which reflect how likely data are to arise from the alternative model, were computed for all analyses with the software JASP using non-parametric Mann–Whitney *U* tests (10,000 samples) and default priors[65]. For between-group comparisons and single-group comparisons against a fixed value: H0: no difference between groups/values and H1: groups/values being different. For (single-group) comparisons against chance level: H0 no difference from chance level (for the tenfold cross-validation, chance level was set to 0.5, and for analyses where values were tested against a permuted distribution, chance level was set to the mean of the distribution) and H1: group being above chance level. Bayes factors were interpreted the following way[37,66]: BF = 1 as no evidence, 1 < BF < 3 as anecdotal evidence, 3 < BF < 10 as moderate evidence, 10 < BF < 30 as strong, and higher values as very strong evidence. Additionally, a Monte Carlo simulation was run using R[37] (see Results section for details; Supplementary Fig. 5).

Neural modelling. The dynamics of macroscopic neural masses shows self-sustained oscillations that can be approximated by a single dynamical variable, namely its phase (θ). Several works show the suitability of this reduction to represent the dynamical behaviour of regions of interest in whole-brain model computational applications[67]. Here, we adopt the following dynamical system[41] to describe the time evolution of auditory and motor phases and their reciprocal influences:

$$\frac{d\theta_i}{dt} = \omega_i + K_{ji}\sin\left(\theta_j(t-\tau) - \theta_i(t)\right) + 0.03\eta(t); \quad i,j = M,A \tag{1}$$

where θ represents the phase of each area, ω is the intrinsic frequency, *K* is the strength of the coupling between areas, η is an additive Gaussian noise centred at zero and *M*/*A* stands for motor/auditory. (Note that equation (1) is used to model motor-to-auditory and auditory-to-motor interactions by switching *i* and *j*.) The auditory–motor transmission delay, τ, is set to 3 ms in line with previous studies[68], and the rest of the parameters (K_{ij} and ω_i) are independently adjusted for each participant and experimental condition. These equations are integrated using the Euler–Maruyama algorithm with a time step of 0.001 s. For each simulation, we calculate the time evolution of the motor and auditory phase during 5 s. The first 3 s are used to thermalize the system (that is, to entrain), while only the last 2 s are submitted to further analyses.

The model described by equation (1) incorporates individual differences through two different parameters: the intrinsic oscillator frequency of each region and the strength of the coupling between oscillators. The auditory intrinsic frequency for each participant and for each simulation was obtained from a Gaussian distribution centred at 4 Hz (σ = 1 Hz), in line with previous work[43]. For the motor frequency, the distribution was narrower (σ = 0.1 Hz) and its centre varies with the experimental condition: 2 Hz for experiment 1 and 4.5 Hz for experiment 2.

Since the oscillators represent motor and auditory cortices and enhanced white-matter pathways connecting these brain regions have been shown in high synchronizers[33], we scaled the strength of the interaction (K_{ij}) with the speech audio–motor synchronization measurement obtained by the SSS-test. This parameter can be disaggregated for each participant *s* as follows:

$$K(s)_{ij} = \alpha_i + \delta_i C(s)_{ij} \tag{2}$$

where α_i is a constant, δ_i is a scale factor and $C(s)_{ij}$ is the individual speech audio–motor synchronization strength obtained with the SSS-test, that is, the PLV normalized between 0 and 1. For both experiments we set $\alpha_{mot} = 0.5$, $\delta_{mot} = 5$ and $K_{AM}(s) = 0.5K(s)_{MA}$.

To simulate the experimental conditions, the time evolution of the motor and auditory phases for each participant was obtained by integrating equation (1) with the corresponding parameters. The synchrony between areas is estimated as the PLV between oscillator phases. With this procedure, we get 59 PLVs for each iteration of experiment 1 and 47 for experiment 2.

Next, we estimated for the high synchronizers the mean phase lag between the simulated auditory activity and the corresponding speech signal. Therefore, we modelled the oscillatory dynamics of the speech envelope (which corresponds to the experimental sinusoid fitting to the envelope of the produced /te/ sequences) as the motor activity with a fixed phase lag. This phase lag represents the time delay between the muscle activity and the speech onset, which has been reported to be jittered by between 100 and 200 ms[45]. Thus, the auditory-to-speech phase lag was defined as $\Delta\theta_{A-Speech} = \langle\theta_M - \theta_A\rangle + \varphi$, with φ assigned for each participant as a random value within $2\pi\omega_M \cdot [0.1;0.2]$.

Finally, we extended the model to simulate a condition in which a rhythmic audio stimulus—in contrast to the stimulus presented in our study—is presented and no motor output is produced (Supplementary Fig. 8). Therefore, we included an external forcing term in the auditory phase oscillator equation as follows:

$$\frac{d\theta_A}{dt} = 2\pi\omega_A + K_{MA}\sin(\theta_M(t-\tau) - \theta_A(t)) + K_E\sin(2\pi\omega_{stim}t - \theta_A(t)) + 0.03\eta(t) \tag{3}$$

where ω_{stim} is the frequency of the external stimulus. In addition, we switch the relative weights of the couplings ($K_{MA}(s) = 0.5K(s)_{AM}$), assuming that during passive listening the auditory-to-motor interaction is stronger than the motor-to-auditory one.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support this study are available from the corresponding authors upon request.

Code availability

Custom code that supports the findings of this study is available from the corresponding authors upon request.

Received: 3 January 2020; Accepted: 9 September 2020;
Published online: 12 October 2020

References

1. Coull, J. T. A in *Brain Mapping* (ed. Toga, A. W.) 565–570 (Academic Press, 2015).
2. Merchant, H. & Yarrow, K. How the motor system both encodes and influences our sense of time. *Time Percept. Action* **8**, 22–27 (2016).
3. Morillon, B. & Baillet, S. Motor origin of temporal predictions in auditory attention. *Proc. Natl Acad. Sci. USA* **114**, E8913 (2017).
4. Morillon, B., Schroeder, C. E. & Wyart, V. Motor contributions to the temporal precision of auditory attention. *Nat. Commun.* **5**, 5255 (2014).
5. Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H. & Lakatos, P. Dynamics of active sensing and perceptual selection. *Cogn. Neurosci.* **20**, 172–176 (2010).
6. Morillon, B., Hackett, T. A., Kajikawa, Y. & Schroeder, C. E. Predictive motor control of sensory dynamics in auditory active sensing. *Curr. Opin. Neurobiol.* **31**, 230–238 (2015).
7. Davis, M. H. & Johnsrude, I. S. Hierarchical processing in spoken language comprehension. *J. Neurosci.* **23**, 3423–3431 (2003).
8. Devlin, J. T. & Aydelott, J. Speech perception: motoric contributions versus the motor theory. *Curr. Biol.* **19**, R198–R200 (2009).
9. Scott, S. K., McGettigan, C. & Eisner, F. A little more conversation, a little less action–candidate roles for the motor cortex in speech perception. *Nat. Rev. Neurosci.* **10**, 295–302 (2009).
10. Wild, C. J. et al. Effortful listening: the processing of degraded speech depends critically on attention. *J. Neurosci.* **32**, 14010–14021 (2012).
11. Chen, J. L., Penhune, V. B. & Zatorre, R. J. Listening to musical rhythms recruits motor regions of the brain. *Cereb. Cortex* **18**, 2844–2854 (2008).
12. Fujioka, T., Trainor, L. J., Large, E. W. & Ross, B. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. *J. Neurosci.* **32**, 1791–1802 (2012).
13. Grahn, J. A. & Rowe, J. B. Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *J. Neurosci.* **29**, 7540–7548 (2009).
14. Besle, J. et al. Tuning of the human neocortex to the temporal dynamics of attended events. *J. Neurosci.* **31**, 3176–3185 (2011).
15. Lakatos, P. et al. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *J. Neurophysiol.* **94**, 1904–1911 (2005).
16. Giraud, A.-L. & Poeppel, D. Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* **15**, 511–517 (2012).
17. Large, E. W. & Jones, M. R. The dynamics of attending: how people track time-varying events. *Psychol. Rev.* **106**, 119–159 (1999).

18. Rimmele, J. M., Morillon, B., Poeppel, D. & Arnal, L. H. Proactive sensing of periodic and aperiodic auditory patterns. *Trends Cogn. Sci.* **22**, 870–882 (2018).
19. Haegens, S. & Zion Golumbic, E. Rhythmic facilitation of sensory processing: a critical review. *Neurosci. Biobehav. Rev.* https://doi.org/10.1016/j.neubiorev.2017.12.002 (2017).
20. 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).
21. Ding, N. et al. Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* https://doi.org/10.1016/j.neubiorev.2017.02.011 (2017).
22. Coupé, C., Oh, Y. M., Dediu, D. & Pellegrino, F. Different languages, similar encoding efficiency: comparable information rates across the human communicative niche. *Sci. Adv.* **5**, eaaw2594 (2019).
23. Morillon, B., Arnal, L. H., Schroeder, C. E. & Keitel, A. Prominence of delta oscillatory rhythms in the motor cortex and their relevance for auditory and speech perception. *Neurosci. Biobehav. Rev.* **107**, 136–142 (2019).
24. Keitel, A., Gross, J. & Kayser, C. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. *PLOS Biol.* **16**, e2004473 (2018).
25. Park, H., Ince, R. A. A., Schyns, P. G., Thut, G. & Gross, J. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr. Biol.* **25**, 1649–1653 (2015).
26. Cason, N., Astesano, C. & Schon, D. Bridging music and speech rhythm: rhythmic priming and audio–motor training affect speech perception. *Acta Psychol. (Amst.)* **155**, 43–50 (2015).
27. Cason, N. & Schon, D. Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia* **50**, 2652–2658 (2012).
28. Falk, S., Lanzilotti, C. & Schon, D. Tuning neural phase entrainment to speech. *J. Cogn. Neurosci.* **29**, 1378–1389 (2017).
29. Hickok, G., Farahbod, H. & Saberi, K. The rhythm of perception: entrainment to acoustic rhythms induces subsequent perceptual oscillation. *Psychol. Sci.* **26**, 1006–1013 (2015).
30. Kösem, A., Basirat, A., Azizi, L. & van Wassenhove, V. High-frequency neural activity predicts word parsing in ambiguous speech streams. *J. Neurophysiol.* **116**, 2497 (2016).
31. Sanabria, D. & Correa, Á. Electrophysiological evidence of temporal preparation driven by rhythms in audition. *Biol. Psychol.* **92**, 98–105 (2013).
32. McPherson, T., Berger, D., Alagapan, S. & Fröhlich, F. Intrinsic rhythmicity predicts synchronization–continuation entrainment performance. *Sci. Rep.* **8**, 11782 (2018).
33. Assaneo, M. F. et al. Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nat. Neurosci.* **22**, 627–632 (2019).
34. Park, H., Thut, G. & Gross, J. Predictive entrainment of natural speech through two fronto-motor top-down channels. *Lang. Cogn. Neurosci.* **35**, 739–751 (2018).
35. Zalta, A., Petkoski, S. & Morillon, B. Natural rhythms of periodic temporal attention. *Nat. Commun.* **11**, 1051 (2020).
36. Assaneo, M. F. & Poeppel, D. The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech-motor rhythm. *Sci. Adv.* **4**, eaao3842 (2018).
37. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: planning for compelling evidence. *Psychon. Bull. Rev.* **25**, 128–142 (2018).
38. Nobre, A. C., Correa, A. & Coull, J. T. The hazards of time. *Sens. Syst.* **17**, 465–470 (2007).
39. Thomas, E. A. C. Reaction-time studies: the anticipation and interaction of responses. *Br. J. Math. Stat. Psychol.* **20**, 1–29 (1967).
40. Grabenhorst, M., Michalareas, G., Maloney, L. T. & Poeppel, D. The anticipation of events in time. *Nat. Commun.* **10**, 5802 (2019).
41. Yeung, M. K. S. & Strogatz, S. H. Time delay in the kuramoto model of coupled oscillators. *Phys. Rev. Lett.* **82**, 648–651 (1999).
42. Poeppel, D. & Assaneo, M. F. Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* https://doi.org/10.1038/s41583-020-0304-4 (2020).
43. Doelling, K. B., Assaneo, M. F., Bevilacqua, D., Pesaran, B. & Poeppel, D. An oscillator model better predicts cortical entrainment to music. *Proc. Natl Acad. Sci. USA* **116**, 10113 (2019).
44. Ruspantini, I. et al. Corticomuscular coherence is tuned to the spontaneous rhythmicity of speech at 2–3 Hz. *J. Neurosci.* **32**, 3786 (2012).
45. Hansen, P., Kringelbach, M., & Salmelin, R. (eds.). *MEG: an Introduction to Methods* (Oxford Univ. Press, 2010).
46. 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).
47. Timm, J., Schönwiesner, M., Schröger, E. & SanMiguel, I. Sensory suppression of brain responses to self-generated sounds is observed with and without the perception of agency. *Cortex* **80**, 5–20 (2016).
48. Cao, L., Thut, G. & Gross, J. The role of brain oscillations in predicting self-generated sounds. *NeuroImage* **147**, 895–903 (2017).
49. Lakatos, P., Gross, J. & Thut, G. A new unifying account of the roles of neuronal entrainment. *Curr. Biol.* **29**, R890–R905 (2019).
50. Chang, E. F., Niziolek, C. A., Knight, R. T., Nagarajan, S. S. & Houde, J. F. Human cortical sensorimotor network underlying feedback control of vocal pitch. *Proc. Natl Acad. Sci. USA* **110**, 2653 (2013).
51. Riecke, L., Formisano, E., Sorger, B., Baskent, D. & Gaudrain, E. Neural entrainment to speech modulates speech intelligibility. *Curr. Biol.* https://doi.org/10.1016/j.cub.2017.11.033 (2017).
52. Riecke, L., Formisano, E., Herrmann, C. S. & Sack, A. T. 4-Hz transcranial alternating current stimulation phase modulates hearing. *Brain Stimulat.* **8**, 777–783 (2015).
53. Henry, M. J. & Obleser, J. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl Acad. Sci. USA* **109**, 20095 (2012).
54. Stefanics, G. et al. Phase entrainment of human delta oscillations can mediate the effects of expectation on reaction speed. *J. Neurosci.* **30**, 13578–13585 (2010).
55. Riecke, L., Sack, A. T. & Schroeder, C. E. Endogenous delta/theta sound-brain phase entrainment accelerates the buildup of auditory streaming. *Curr. Biol.* **25**, 3196–3201 (2015).
56. Zoefel, B., Davis, M. H., Valente, G. & Riecke, L. How to test for phasic modulation of neural and behavioural responses. *NeuroImage* **202**, 116175 (2019).
57. Repp, B. H. & Penel, A. Rhythmic movement is attracted more strongly to auditory than to visual rhythms. *Psychol. Res.* **68**, 252–270 (2004).
58. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
59. Boersma, P. Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2001).
60. Kontsevich, L. L. & Tyler, C. W. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Res.* **39**, 2729–2737 (1999).
61. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118**, 887–906 (2005).
62. *NSL MATLAB Toolbox* (Neural Systems Laboratory, University of Maryland, 2003).
63. Moskowitz, C. S. & Pepe, M. S. Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostat. Oxf. Engl.* **5**, 113–127 (2004).
64. Huang, Y., Sullivan Pepe, M. & Feng, Z. Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–1188 (2007).
65. *JASP version 0.12* (JASP Team, 2020).
66. Lee, M. D. & Wagenmakers, E.-J. *Bayesian Cognitive Modeling: A Practical Course* pp. xiii, 264 (Cambridge Univ. Press, 2013).
67. Cabral, J. et al. Exploring mechanisms of spontaneous functional connectivity in MEG: how delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations. *NeuroImage* **90**, 423–435 (2014).
68. Guenther, F. H., Ghosh, S. S. & Tourville, J. A. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain Lang.* **96**, 280–301 (2006).

Acknowledgements

We thank M. Grabenhorst, J.-R. King and L. Gwilliams for their valuable input regarding the data analysis, M. Fichter for data recordings and S. Brendecke for graphics support. This work was funded by the Max-Planck-Institute for Empirical Aesthetics. The funders had no role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

Author contributions

M.F.A., J.M.R. and D.P. conceived of and designed the experiments. J.M.R. collected the data. M.F.A. and J.M.R. conceived and designed the analyses. M.F.A. performed the main analyses and contributed the SSS-test analysis toolbox. Y.S.P. and M.F.A. generated the computational model. M.F.A., J.M.R. and D.P. interpreted the results. M.F.A. and J.M.R. wrote the manuscript. All authors edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00962-0.

Correspondence and requests for materials should be addressed to M.F.A. or J.M.R.

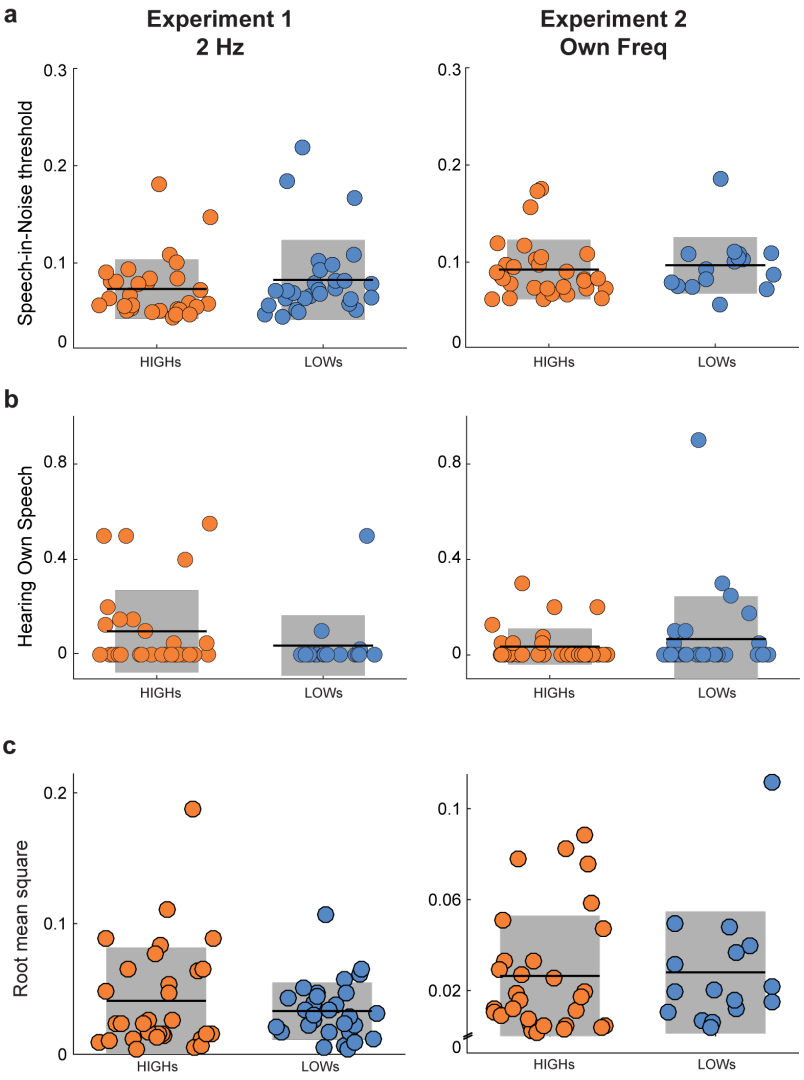
Peer review information Primary Handling Editor: Marike Schiffer.

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

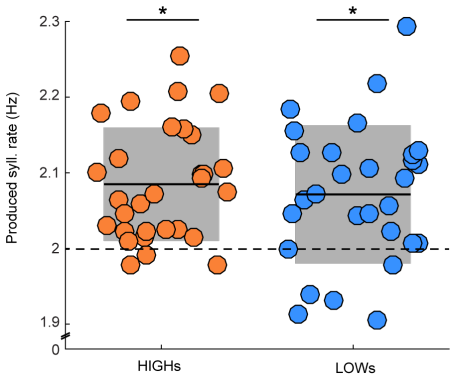
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

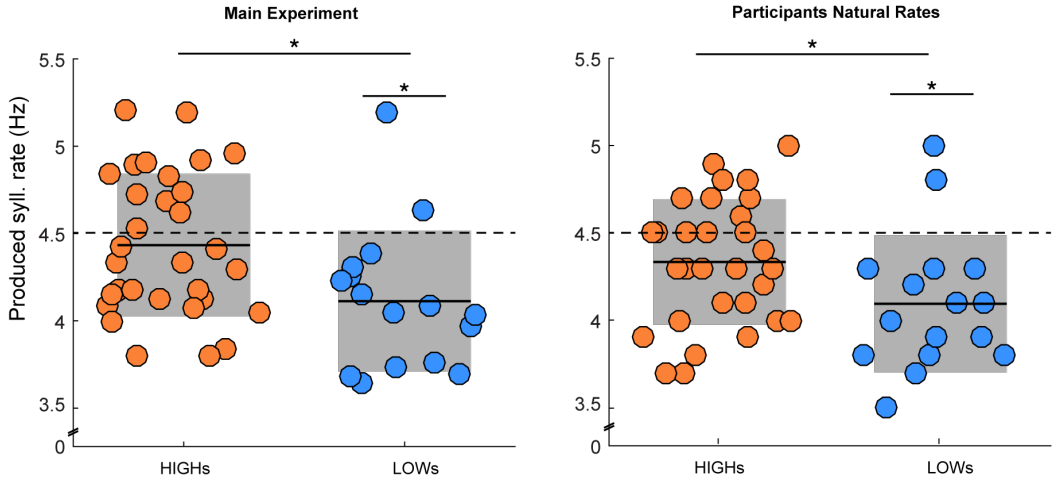
Supplementary Information



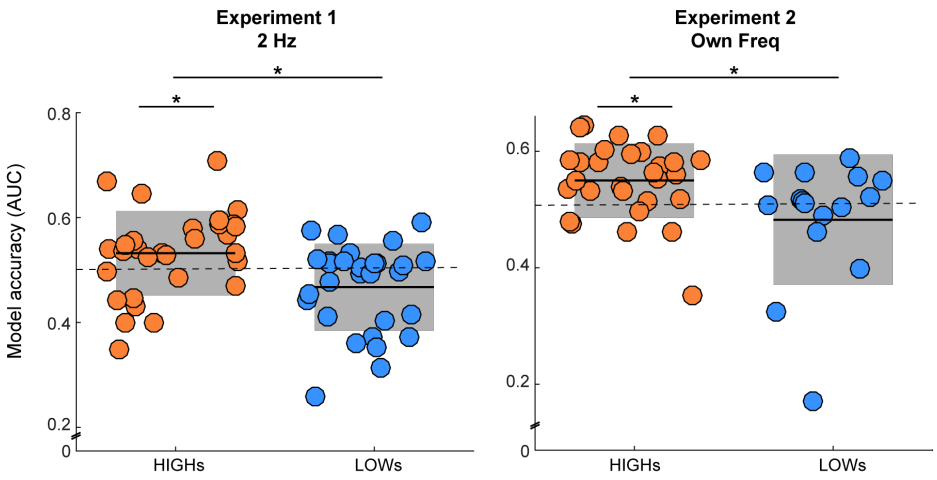
Supplementary Figure 1 Non relevant parameters. There were no significant differences between the high and low synchronizers in any of these measurements. **a,** Speech-in-Noise (SiN) thresholds (Experiment 1: W=503.0, p=0.309, r=0.156, CI=[-0.139, 0.426], BF_{01}=2.31; $mean_{highs}$=0.073, $mean_{lows}$=0.083, n_{highs}=30, n_{lows}=29; Experiment 2: W=286, p=0.298, r=0.192, CI=[-0.159 0.499], BF_{01}=2.51, $mean_{highs}$=0.092, $mean_{lows}$=0.097, n_{highs}=30, n_{lows}=16; **b,** Reported own speech perception (Experiment 1: W=445.5, p=0.846, r=0.024, CI=[-0.266 0.311], BF_{01}=3.21; $mean_{highs}$=3.50%, $mean_{lows}$=6.64%, n_{high}s=30, n_{lows}=29; Experiment 2: W=173, p=0.136, r=-0.228, CI=[-0.530 0.126], BF_{01}=1.97, $mean_{highs}$=9.91%, $mean_{lows}$=3.91%, n_{highs}=28, n_{low}s=16; **c,** Loudness of the produced speech, root mean square of the produced signals (Experiment 1: W=457, p=0.918, r=9.017, CI=[-0.271 0.302], BF_{01}=3.64; $mean_{highs}$=0.041, $mean_{lows}$=0.033, n_{high}s=31, n_{lows}=29; Experiment 2: W=263, p=0.486, r=0.131, CI=[-0.224 0.456], BF_{01}=2.61, $means_{highs}$=0.055, $mean_{lows}$=0.062, n_{highs}=31, n_{lows}=15). In all panels: dots represent individual subjects, orange/blue indicates high/low synchronizers; black line the mean value and shadowed the standard deviation.



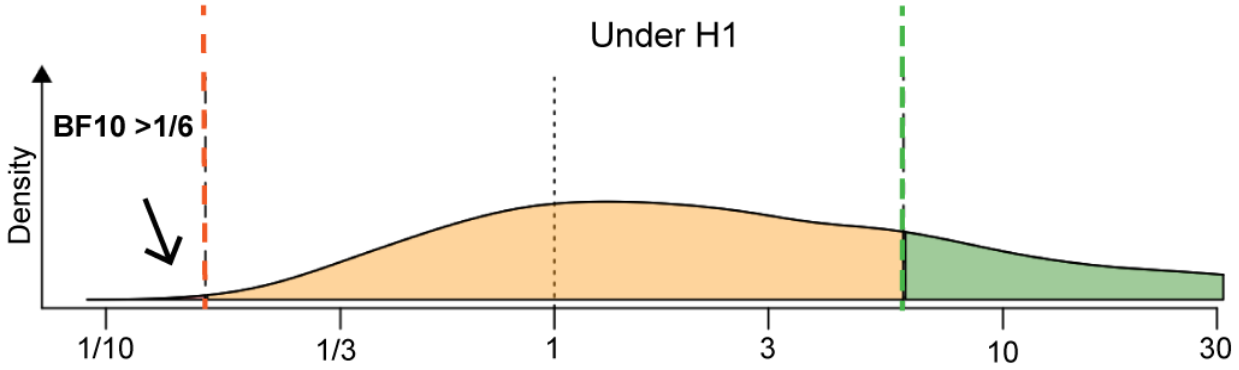
Supplementary Figure 2 Speech production rate across trials, Experiment 1. In Experiment 1 the produced rate was significantly faster than the target rate (2 Hz, dashed line) for both groups (n_{highs}=30, n_{lows}=29, highs: V=453, p_{vs2Hz}<0.001, r=0.948, CI=[0.887 0.977], BF_{10}=2824, mean=2.09, lows: V=377, p_{vs2Hz}<0.001, r=0.733, CI=[0.477 0.875], BF_{10}=555.5, mean=2.07). Production rates did not differ between groups (W=408, p=0.69, r=-0.062, CI=[-0.345 0.231], BF_{01}=3.17). In all panels: dots represent individual subjects, orange/blue indicates high/low synchronizers; black line the mean value and shadowed the standard deviation.



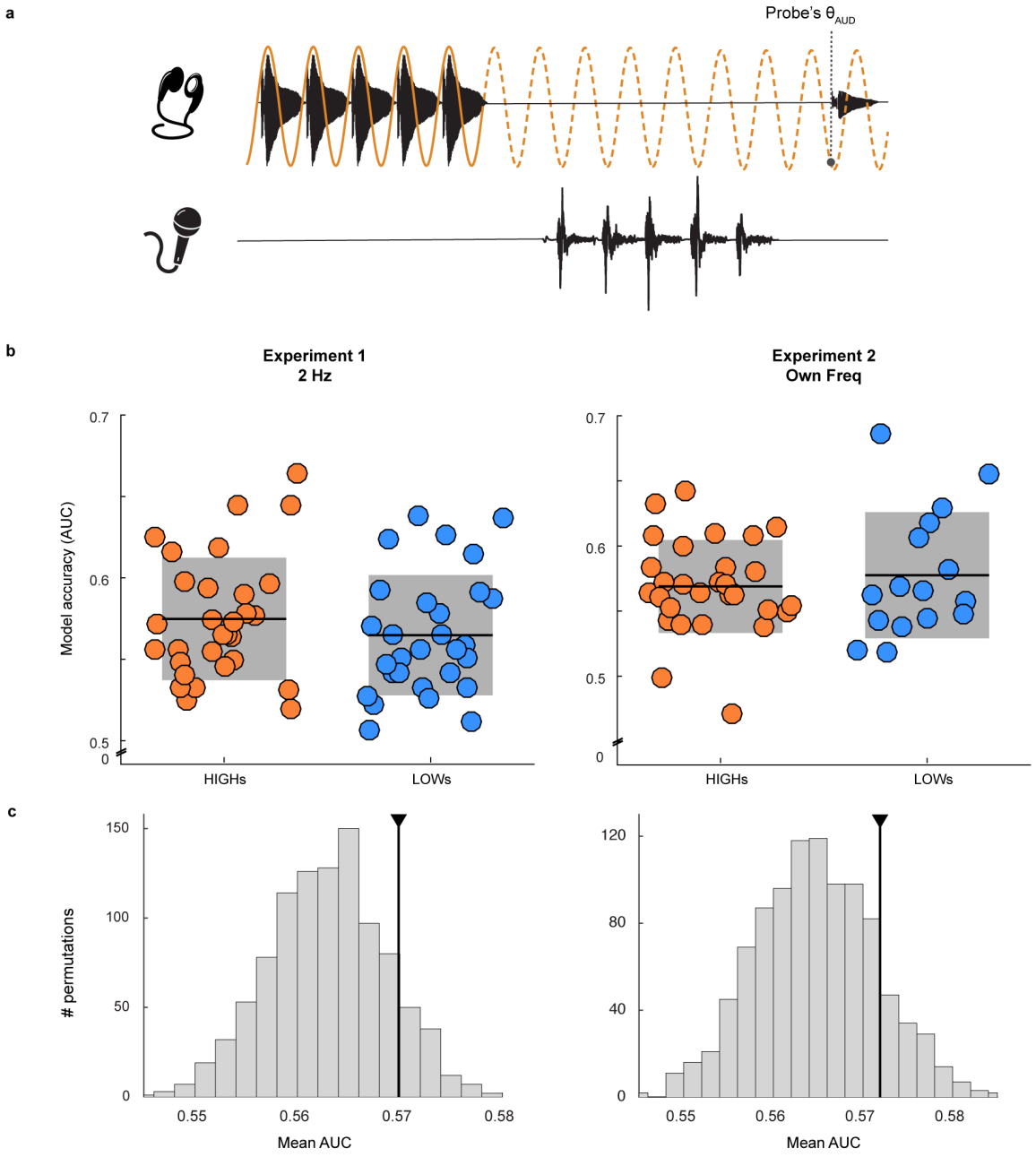
Supplementary Figure 3 Speech production rates, Experiment 2. Left panel: produced rate during the main experiment. Right panel: Target frequency extracted prior to the experiment, the natural production rate that was extracted for each participant and used as the target rate for the main experiment. In both panels only the production rates of the lows are significantly different than the primed rate (4.5 Hz, dashed line; Main Experiment: highs: V=199, $p_{vs4.5Hz}$ = 0.497, r=-0.144, CI=[-0.504 0.259], lows: V=13, $p_{vs4.5Hz}$ = 0.003, r=-0.809, CI=[-0.933 -0.513]; Target frequencies: highs $p_{vs4.5Hz}$ = 0.11, lows: $p_{vs4.5Hz}$ = 0.004) and lows have significantly lower rates compared to the highs (Main Experiment: W=130, p=0.012, r=-0.458, CI=[-0.691 -0.140], BF_{10}=3.18; $highs_{mean}$=4.45, $mean_{lows}$=4.11; n_{highs}=30, n_{lows}=16. Target frequencies: W=148.5, p=0.035, r=-0.381, CI=[-0.638 -0.047], BF_{10}=1.02, $mean_{highs}$=4.36 Hz, $mean_{lows}$=4.09 Hz; n_{highs}=30, n_{lows}=16). In all panels: dots represent individual subjects, orange/blue indicates high/low synchronizers, black line the mean value, asterisks p < 0.05 and shadowed the standard deviation.



Supplementary Figure 4 Speech production significantly entrains speech perception, in high synchronizers. Model accuracy was computed applying a repeated 10-fold cross validation method. Each trial within participants was randomly assigned to one of 10 groups, accuracy was tested on each of the 10 groups while the model was trained on the other nine. The obtained 10 AUCs were averaged to get one output value. The whole procedure was repeated 10 times. Results were averaged across repetitions and within subject. Accuracy was enhanced for the highs compared to the lows in both experiments (Experiment 1, left column: W=249, p = 0.007, r=-0.407, CI=[-0.624 -0.132], BF_{10}=4.82, n_{highs}=30, n_{lows}=28; Experiment 2, right column: W=107, p = 0.005, r=-0.508, CI=[-0.728 -0.194], BF_{10}=3.54, n_{highs}=29, n_{lows}=15). Furthermore, only for high synchronizers the models performed above chance level in both experiments (Experiment 1: highs $p_{vs0.5}$ = 0.041, BF_{+0}=2.6, lows: $p_{vs0.5}$ = 0.18, BF_{0+}=11.49; Experiment 2: highs $p_{vs0.5}$ = 0.0014, BF_{+0}=80.56, lows: $p_{vs0.5}$ = 0.83, BF_{0+}=6.01). In all panels: dots represent individual subjects, orange/blue indicates high/low synchronizers, black line the mean value, asterisks p < 0.05 and shadowed the standard deviation.

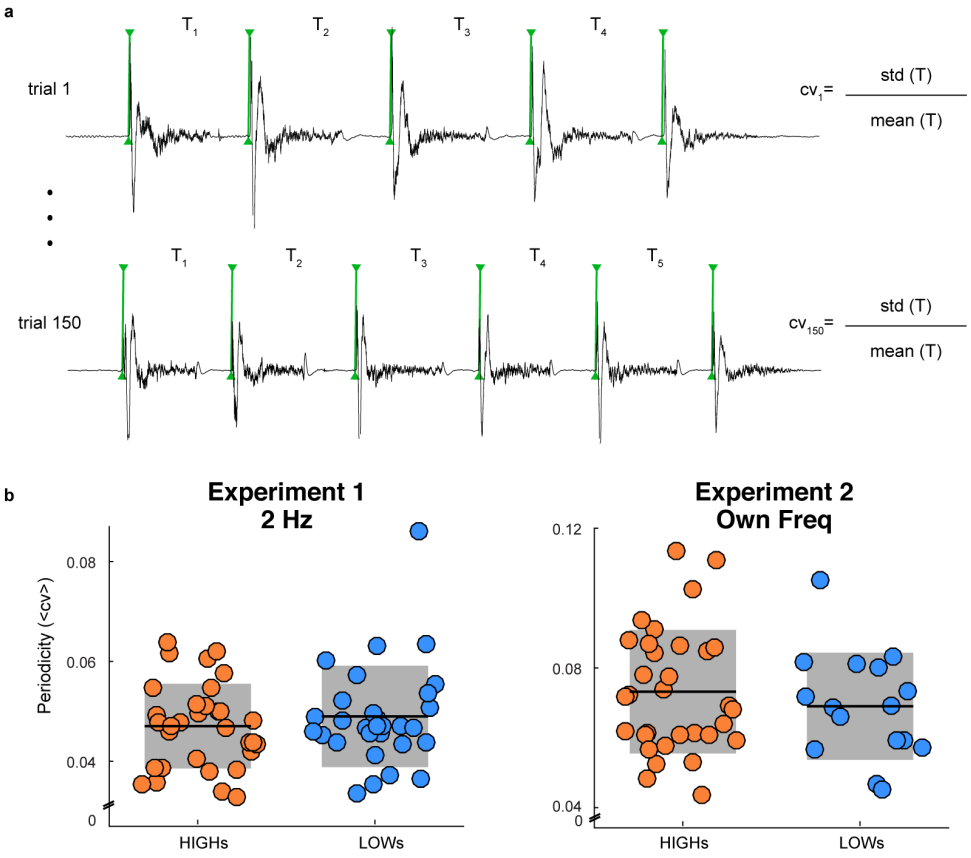


Supplementary Figure 5. Monte Carlo Simulation estimating the probability of false evidence for H0 given H1. The distribution of bayes factors obtained in a simulation[38] (10.000 iterations; Cauchy prior = √2/2; with a sample size of N =16 per iteration) is displayed. Under the assumption that H1 is true given the moderate effect size we observed (in Exp. 2 for highs: r=0.656) the probability of false evidence (highlighted in red) for H0 with a moderate bayes factor (BF_{10} =1/6; red dashed line) is 0.1% and thus very low. The findings support our interpretation that the bayes factor observed in our study (arrow) unlikely indicates false evidence for H1. The probability of observing evidence for H1 is 29.1% highlighted in green, and the probability of inconclusive evidence (0.1667 < BF_{10} < 6) is 70.8% highlighted in orange.

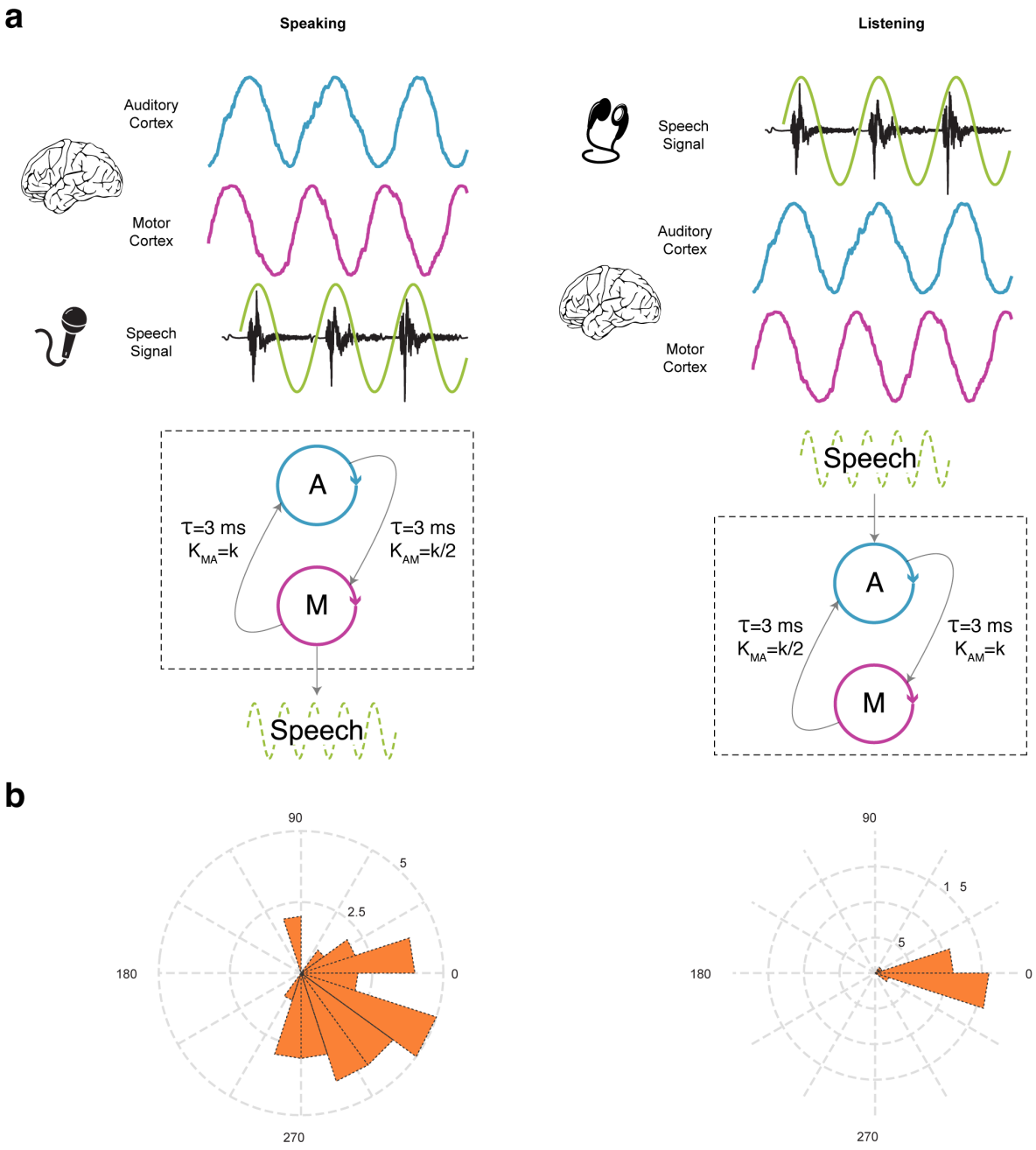


Supplementary Figure 6 Auditory to auditory entrainment. a, Models were fitted using the auditory phase at which the probe was presented as a predictor of the participant's answer. The auditory phase was computed by extending the oscillation defined by the rhythmic target /te/ sequence (brown line). The headset and microphone icons are used from Stock unlimited LCC. **b,** Model accuracies were not different between low and high synchronizers in neither of the experiments (Experiment 1, left column: W=367, p = 0.309, r=-0.156, CI=[-0.426 0.136]; BF_{01} = 2.111, n_{highs}=,30 n_{lows}=29; Experiment 2, right column: W=237, p = 0.955, r=-0.012, CI=[-0.351 0.329], BF_{01} = 3.18, n_{highs}=30, n_{lows}=16); **c,** Histograms of 1000 mean AUC values obtained by different iterations of permuted answers within subject. Black lines indicate the non-permuted mean AUC. The non-permuted mean AUC remained below the 95th percentile computed on the Null-distribution ($mean_{nullExp1}$= 0.5629, $mean_{nullExp2}$=0.5647) for both experiments (Experiment 1,

left column: mean AUC = 0.569 > 95_{prctl} = 0.573, BF_{0+} = 1.77; Experiment 2: mean AUC = 0.571 > 95_{prctl} = 0.576, BF_{0+} = 2.6).



Supplementary Figure 7 Periodicity of the produced sequences. a, For each participant we estimated the periodicity of the produced /te/ sequences by mean of the following sequential steps: 1. by visual inspection we obtained the onset of each produced /te/ (green lines), 2. for every trial we calculated the coefficient of variance (cvi) of the time intervals between successive syllable onsets (T) and 3. the coefficient of variance was averaged across trials (<cv>). **b,** Periodicity displayed no significant difference between groups in neither of the experiments. Right panel: Experiment 1, W=510, p=0.680, r=0.063, CI=[-0.223 0.39], $mean_{highs}$=0.047, $mean_{lows}$=0.049, BF_{01} = 3.38, n_{highs}=32, n_{lows}=30. Left panel: Experiment 2, W=216, p=0.392, r=-0.156, CI=[-0.468 0.190], BF_{01}= 2.57, n_{highs}=32, n_{lows}=16. In all panels: dots represent individual subjects; orange/blue high/low synchronizers; black line the mean value and shadowed the standard deviation.



Supplementary Figure 8. With an external rhythmic stimulus the model predicts a tight alignment of the phase lags between cortices across participants. a, Schematic representation of the model: On the left, speech production; on the right speech perception. Dashed box indicates that in this case the phase locking value was estimated between auditory and motor cortices (blue and magenta traces, respectively). The headset and microphone icons are used from Stock unlimited LCC. The brain icon is used from Pixabay. The A/M interaction schematic is adapted from[36]. **b,** Phase lag between auditory and motor oscillators for high synchronizers obtained by the two different models: speaking, on the left, and listening on the right (see Methods). Parameters were set at the conditions of Experiment 2.