Published in final edited form as:
Nat Neurosci. 2019 April ; 22(4): 627–632. doi:10.1038/s41593-019-0353-z.

Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning

M. Florencia Assaneo[1,*], **Pablo Ripolles**[1,*], **Joan Orpella**[2,3,4,*], **Wy Ming Lin**[1], **Ruth de Diego-Balaguer**[2,3,4,5,‡], and **David Poeppel**[1,6,‡]

[1]Department of Psychology, New York University, New York, New York, 1003, USA.

[2]Cognition and Brain Plasticity Unit, IDIBELL, L'Hospitalet de Ll., 08907, Spain

[3]Department of Cognition, Development and Educational Psychology, University of Barcelona, Barcelona, 08035, Spain

[4]Institute of Neuroscience, University of Barcelona, Barcelona, 08035, Spain

[5]ICREA, Barcelona, 08010, Spain

[6]Neuroscience Department, Max-Planck Institute for Empirical Aesthetics, Frankfurt, 60322, Germany.

Abstract

We introduce a deceptively simple behavioral task that robustly identifies two qualitatively different groups within the general population. When presented with an isochronous train of random syllables, some listeners are compelled to align their own concurrent syllable production with the perceived rate, while others remain impervious to the external rhythm. Using both neurophysiological and structural imaging approaches, we show group differences with clear consequences for speech processing and language learning. When listening passively to speech, high synchronizers show increased brain-to-stimulus synchronization over frontal areas and this localized pattern correlates with precise microstructural differences in the white matter pathway connecting frontal to auditory regions. Finally, the data expose a mechanism that underpins performance on an ecologically relevant word-learning task. We suggest that this task will helps to better understand and characterize individual performance in speech processing and language learning.

[*,‡]These authors contributed equally to this work

Author contributions:
MFA, PR, JO, WML, RDB & DP designed the research and wrote the paper; MFA, PR, JO, and WML acquired and analyzed the data.

Competing interests
The authors declare no competing interests.

Supplementary Materials:
Figures S1–S9
Table S1

INTRODUCTION

The ability to synchronize a motor output to an auditory input – i.e. tapping or dancing to music with a groove – is a basic trait present in humans from birth[1], with important cognitive implications[2,3]. From a phylogenetic perspective, spontaneous synchronization (i.e. without explicit training) to an external rhythm is argued to be a unique characteristic of vocal learning species, including humans[4]. The study of this distinctive attribute has typically focused on how body movements are entrained by non-speech signals - e.g. music[5] or a beat[6]. One foundational question, in this context, however, has not been systematically investigated: Do humans *spontaneously* align their speech motor output to auditory speech input? Resolving the role of audio-motor synchronization[7–9] in the context of speech processing is a critical step for the characterization of the complex functional and structural neural architecture of language.

To address these questions, we designed a simple behavioral protocol to explore the spontaneous synchronization of speech (SSS-test). The results show, contrary to the previous literature, that a substantial part of the population does not show *speech-to-speech* synchronization. Thus, we further explored the functional and structural brain correlates associated with the absence of synchrony. Finally, we turn to the key issue of whether the behavioral findings on audio-motor synchronization and their neural substrate extend to tasks that address more typical questions of speech processing and language learning.

RESULTS

Spontaneous speech synchronization reveals a bimodal distribution

Participants (N = 84) completed two behavioral blocks of listening to a rhythmic train of syllables – at 4.5 syllables-per-second, the characteristic speech rate across languages[10,11] – lasting one minute each, while concurrently whispering the syllable *“tah”* (Fig. 1A). At the end of each block, participants indicated whether a given target syllable was presented. Crucially, participants were instructed to correctly recall the syllables; there was no explicit instruction to synchronize the utterances to the external rate. We first examined the degree of synchronization between the produced utterances and the input signal by computing the phase locking value (PLV) between their envelopes around the given syllable rate (4.5 Hz, Fig. 1A). Surprisingly, participants’ PLVs yield a bimodal distribution (Fig. 1B), suggesting a segregation of our cohort into distinct populations of high and low synchronizers (N_{high} = 43 and N_{low} = 41). No difference was found between groups in terms of language background, age, or gender. However, the high synchronizers showed, overall, more years of musical training than the lows (Mann-Whitney-Wilcoxon test, two-sided p =0.0033; Fig. S1). Still, musical experience by its own did not segregate the population into two groups (Fig S1F).

Next, we analyzed the distribution of the phase differences between perceived and produced syllables for the high synchronizers. We found a non-uniform distribution (N_{high} = 43, Rayleigh Test, two-sided p < 0.001; Fig 1C) with the phase lags concentrated around zero (95% CI [−0.21 0.26]): the highs adjusted their production to be in phase with the perceived syllables. The average spectra of the produced speech envelopes also exhibited striking group differences. The high synchronizers showed a pronounced peak at the given syllable rate, indicating high stability in maintaining the rhythm. The low synchrony group was less stable, exhibiting a broader peak (Fig. 1D). Two additional experiments were conducted to further assess empirically whether the synchrony of the produced utterances was indeed driven by an *interaction* between perceived and produced speech rhythms (rather than related to the ability to maintain a tempo). First, a subset of the participants ($N_{high} = 13$, $N_{low} = 12$) completed an additional block of whispering "*tah*" while listening to white noise (no-rhythm condition). Speech rhythmicity was strongly reduced in high synchronizers during the no-rhythm condition, relative to the rhythmic one, and stayed unchanged in low synchronizers (Fig. 1E). Second, a new cohort of participants completed a modified version of the SSS-test in which the rate of the perceived syllables was gradually increased (see Methods). Crucially, the distribution also displayed two peaks under this condition (N = 55, Fig. S2). High synchronizers adjusted their speech output to (multiple) changes in the tempo of the perceived speech (see Fig. S2B, top panel). This result provides additional compelling evidence that a subgroup of the population adapted the produced rhythm to the perceived one. Furthermore, in most cases, participants were not aware of the rate's shift (Fig S2C), highlighting the unconscious and automatic nature of the phenomenon. These combined findings support the conjecture that participants exhibited two qualitatively different behaviors: whereas the speech output of high synchronizers was entrained by the external auditory speech rate, low synchronizers showed no interaction between the produced and the perceived speech rhythm.

We next determined that subjects' degree of audio-motor synchrony is stable over time, suggesting that synchronization-type is a consistent individual trait. This is demonstrated by the highly correlated PLV scores, both across blocks within session (N=84, Spearman correlation coefficient r = 0.86, p < 0.001; Fig. 1F) as well as across sessions distant in time (34 participants repeated the test one month later; Spearman correlation coefficient r = 0.78, p < 0.001; Fig. 1G). This stability over time proves the SSS-test's reliability in clustering participants into high and low synchrony groups via a straightforward measure of their phase locking to a regular pacing signal.

To further verify that the phenomenon is robust and replicable, we developed and conducted an online version of the SSS-test both for the stable and the accelerated syllable rate using the Amazon Mechanical Turk platform. These two additional replications (N = 144 for the stable, N = 60 for the accelerated version) underscore the reliability of the bimodal distribution under less controlled conditions (Fig. S3A for stable rate and S4A for accelerated) and also allowed us to explore differences between groups in their perception/production abilities. From the perceptual standpoint, the high synchronizers who completed the online version of the stable SSS-test (N=144), were marginally better than the lows (Mann-Whitney-Wilcoxon test, two-sided p = 0.093; Fig. S3B). From the production one, the low synchronizers were marginally less accurate (not corrected for multiple comparisons) in keeping a precise syllable rate (4.5 syll. per sec) without feedback (on white noise): they tended to produce lower syllable rates (Fig. S3C). While these results imply that the two populations - identified by our one-minute test relying on speech audio-motor interactions - also differ, however modestly, at other levels of perception/production abilities, further work is needed to delineate the extent of the predictive power of our test with regard to other capabilities.

In light of the reliability of our findings, we next pursued the following question: Does the clear grouping based on the straightforward behavioral paradigm reflect neural properties and behavioral consequences that have broader significance? These behavioral data invite the hypothesis that functional and/or structural brain differences underlie the high versus low segregation. To address this question, we next acquired both neurophysiological and structural data from a subgroup from the original cohort ($N_{high} = 18$, $N_{low} = 19$; Fig. 1E).

Neural distinction between groups: neurophysiological data

In a magnetoencephalography (MEG) experiment, participants listened to rhythmic trains of syllables (4.5 syllables per second), now *passively* (i.e., no whispering). They were instructed to listen attentively to the syllables and to indicate after each stream whether a given set of syllables had been presented. Subjects' attention on the syllables could be verified, since they performed the task above chance level (N=37, Wilcoxon Signed-Rank test, two-sided $p = 0.011$; Fig. S5), although there was no significant difference in the performance between groups ($N_{high} = 18$, $N_{low} = 19$, Mann-Whitney-Wilcoxon, two-sided $p = 0.96$). Caution is required in interpreting this behavioral result since syllable recognition is rather poor for the entire participant cohort. This derives from the fact that we designed an extremely difficult task (12 synthesized syllables, coarticulated for 2 minutes) to maximize participants' attention during the two minutes of the Syllable Perception Task. We then computed the PLV between elicited brain activity and the envelope of the auditory stimuli (i.e. brain-to-stimulus synchrony; Fig. 2A) in the frequency band corresponding to the perceived syllable rate (4.5 +/− 0.5 Hz). Given that the group segregation relies on a speech audio-motor task, we centered our analyses in bilateral frontal as well as temporal regions implicated in speech production and perception, respectively. A first comparison between groups restricted to frontal regions-of-interest (ROIs) revealed that high synchronizers showed enhanced brain-to-stimulus synchrony in left inferior and middle frontal gyri, more precisely in left Brodmann Areas 44, 45, 46 and 9 ($N_{high} = 18$, $N_{low} = 19$, Mann-Whitney-Wilcoxon test, two-sided $p < 0.05$, FDR-corrected; Fig S6A and Fig. 2B). Interestingly, previous data show that, during overt speech production, the control of temporal speech patterns is likely governed by these regions[12,13]. Thus, our results suggest that areas related to speech timing during *production* are also implicated during speech *perception* to track the perceived syllable rate (note that no motor production occurred during the MEG session).

In contrast, the same analysis performed on temporal ROIs yielded no significant differences between the groups (Fig. S6). However, the asymmetry of the entrainment in early auditory regions was significantly different between groups (see Methods and Fig. S7).

Structural distinction between groups: anatomic connectivity data

Having observed neurophysiological group differences, we then acquired diffusion weighted MRI data (DW-MRI) from the same cohort to quantify potential differences in the white matter pathways connecting the frontal and auditory regions that distinguished the groups in terms of brain-to-stimulus synchrony. Excitingly, high synchronizers showed enhanced microstructural properties in the white matter neighboring the auditory cortex (Fig. 3A). Specifically, we found a distinct lateralization pattern in a white matter cluster (N_{high} = 18, N_{low} = 18, FWE-corrected at the peak voxel, two-sided p = 0.024; Fig. 3B) likely part of the arcuate fasciculus[14–16] that differentiated between groups, with high synchronizers showing significantly greater left lateralization. Note that no significant clusters were obtained in ventral white matter pathways connecting frontal and temporo-occipital regions (see Methods). Crucially, this structural difference relates to both the auditory (Fig. S7B) and frontal (Fig. 3C) neurophysiological results: increased leftward lateralization in the white matter was related to higher brain-to-stimulus synchrony in left frontal regions (Fig. 3C) and to more symmetrical auditory entrainment (Fig. S7C). Virtual dissections (i.e., tractography) further showed that the volume of the left arcuate (but not of the left inferior longitudinal or inferior fronto-occipital fasciculi, which were also dissected as a control) not only differentiated between groups, but was also related to left frontal neurophysiological brain-to-stimulus synchrony (see Methods and Fig. S8).

Spontaneous Speech Synchronization test predicts word learning

Previous research documents that the early stages of word learning capitalize on the interaction between auditory and frontal regions and the white matter pathways connecting them[15]. To test for a principled link between these observations and our simple behavioral test, a new sample of participants (N = 44) was recruited to complete the SSS-test as well as a word-form learning task. More precisely, since we expected that speech synchronization should most clearly benefit segmentation abilities, a classical statistical learning paradigm was chosen[17,18]. In this paradigm, participants listened for two minutes to a continuous repetition of four tri-syllabic pseudo-words, which were randomly concatenated without silence gaps between them. Next, they completed a testing phase to assess whether they correctly segmented the pseudo-words. The histogram of the PLVs, obtained with the SSS-test for this smaller group, also displayed two peaks, replicating the bimodal distribution shown by the original cohort (Fig. 4A). By splitting this new population into high and low synchronizers (using the median PLV of the first cohort, Fig. 4A), we find that high synchronizers have a significant learning advantage in the phonological word-learning task (N_{high} =24, N_{low}=20; r = 0.4, Rank-Biserial Correlation; Mann-Whitney-Wilcoxon test, two-sided p=0.024; Fig. 4B). This learning benefit was also replicated in the additional cohort of participants who completed an online version of the word learning task in addition to the online and accelerated version of the test (N_{high} =25, N_{low}=35; r = 0.37, Rank-Biserial Correlation; Mann-Whitney-Wilcoxon test, two-sided p=0.015; see Fig. S4B). We hypothesize that: (i) in line with previous work[8,19], the increment of synchronization in the frontal region, enhanced in the high synchronizers, facilitates the parsing of the syllables by aligning attention to their onsets; (ii) better parsing improves the extraction of the statistical relations between the syllables; (iii) likewise, the prediction of one syllable following the other helps to create a better phonological trace of the whole word. In conclusion, enhanced audio-motor interaction as measured with our approach (the SSS-test) is not only reflected in functional and structural properties of frontal and temporal areas but has compelling consequences for language learning.

DISCUSSION

The combined behavioral, neurophysiological, and neuroanatomic results reveal a fundamental phenomenon: whereas some individuals are compelled to spontaneously align their speech output to the speech input, others remained impervious to the external rhythm (see Fig. S9 for a depiction of the joint bimodal distribution of all our experiments using the SSS-test). We speculate that such distinct populations of high and low synchronizers emerge from the *spontaneous* nature of the synchrony induced by the SSS-test (the goal of the task being orthogonal to the synchronization). This contrasts with previous research showing more homogenous entrainment patterns when synchronization to an external auditory signal is explicitly requested[2,20]. The behavioral pattern we have discovered correlates with neurophysiological and structural features within key elements of the speech brain network[21–23]: production areas (IFG), perception areas (early auditory cortex), and the white matter connecting them (see tractography analysis in Fig S8)[24]. Excitingly, the fact that our results scale up to an ecologically relevant task[18], word learning in the context of speech segmentation, has theoretical and practical consequences for how we understand and study individual differences in cognition and learning[25,26].

Our ability to speak relies on a widely distributed and highly interconnected audio-motor network[27,28]. We hypothesize that an interplay between structural and physiological predispositions (roughly, nature) and experience-dependent tuning (roughly, nurture) can generate moderate modifications to the components of the speech audio-motor network that, due to its coarse connection[29], result in large consequences at the functional and behavioral level. Specifically, a subtle enhancement in the structure of the white matter connecting auditory and motor regions could improve the synchronization (flow of information[30]) between temporal and frontal areas, in turn eliciting the effects observed in these experiments. Previous research shows that white matter located in the same region as the cluster highlighted by our study undergoes microstructural changes through musical training[9,31]. In line with these studies, we found that the high synchronizers have, overall, more years of musical training than the lows. However, in our work, musical training on its own does not follow a bimodal distribution, suggesting that musical experience is one of many factors defining group membership to high or low synchronizers.

In summary, we show a deceptively simple behavioral task (the SSS-test) capitalizing on individual differences that turn out to be predictive of audio-motor synchronization, neurophysiological function, brain anatomy, and performance on an ecologically relevant word-learning task. The use of such a test can help to better characterize individual performance, leading to new discoveries related to speech processing and language learning that could have been masked by pooling together populations with substantially different neural and behavioral attributes.

ONLINE METHODS

Participants

A first cohort of 84 participants initially completed the SSS-test (32 males; mean age, 28; age range, 19 to 55). From this, a subgroup of 37 subjects (right handed; 18 males; mean age, 30; age range, 21 to 55) also underwent the MEG and DW-MRI protocols. The original subgroup comprised four additional participants, but these had to be removed due to artifactual MEG (3 participants) or DW-MRI data (1 participant). The MEG session took place at least 4 days after the DW-MRI session. Both protocols were completed within one month after the SSS-test.

A second cohort of 44 individuals (11 males; mean age, 21; age range, 19 to 31) completed the SSS-test and the word-learning task.

A third cohort of 62 participants completed the accelerated version of the SSS-test. Seven participants were removed, because they spoke loudly instead of whispering or they stopped whispering for time periods longer than 4 sec. The data from 55 participants (19 males; mean age, 23; age range, 18 to 36) were analyzed.

Two extra cohorts, one of 200 and one of 100 participants, completed the online version of the regular and the accelerated SSS-test, respectively. 56 participants from the regular group, and 40 from the accelerated were removed for non-optimal conditions in their recordings (noisy recording, participant did not use headphones, participant spoke loudly instead of whispering, or stopped whispering for time periods longer than 4 sec). The final number of participants submitted to the analyses were: 144 (80 males; mean age, 34; age range, 19 to 55) for the regular SSS-test; and 60 (37 males; mean age, 35; age range, 19 to 51) for the accelerated SSS-test.

All participants were native English speakers with self-reported normal hearing and no neurological deficits. They were paid for taking part in the study and provided written informed consent. All protocols were approved by the local Institutional Review Board (New York University’s Committee on Activities Involving Human Subjects).

No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications[7,15,17].

Statistical Analyses

Data distribution was not formally tested. Instead, we used non-parametric Mann-Whitney-Wilcoxon and Wilcoxon signed-rank tests were used for between and within subject comparisons, respectively. Multiple comparisons were controlled using a False-Discovery Rate correction (the only exception is the tract-based spatial statistics white matter analyses which use a FWE correction based on threshold-free cluster enhancement and a nonparametric permutation test). Non-parametric Spearman’s rank correlations were used to assess the relationship between variables. In addition, we used the Robust Correlation Toolbox[32] to ensure the robustness of the relationship between the structural and neurophysiological data. In particular, we used Spearman skipped correlations[33,34] with percentile bootstrap 95% confidence intervals (CIs; calculated by resampling pairs of observations) for each correlation. Skipped correlations involve multivariate outlier detection and provide a more robust measure of correlation[35]. Bootstrap confidence intervals provide an additional way to test whether two variables are truly correlated, as if the CIs include 0, the null hypothesis cannot be rejected[32].

Effect sizes were calculated using Rank-Biserial correlations which can be read as a Spearman's correlation coefficient[36], or as the difference between the proportion of favorable and unfavorable evidence[37]. In our study, all effect sizes are above r = 0.36. An *r-value* of 0.36 means that the favorable evidence outweighs the unfavorable 68% to 32%.

Data collection and analysis were performed blind to the conditions of the experiments for the behavioral tests but not for the MEG and structural analyses, as subjects had already been divided into high and low synchronizers.

Stimulus presentations order was randomized for all experiments with more than one stimulus.

Phase locking value

Throughout the study, the synchronization between two signals was measured by the phase locking value (PLV) between them. The PLV was computed using the following formula:

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

where t is the discretized time, T is the total number of time points and θ_1 and θ_2 the phase of the first and the second signals, respectively.

Stimuli

Five sets of syllables (G1 to G5) were created to be used in the MEG task. Each set consisted of 12 distinct syllables (unique consonant-vowel combinations) handpicked to maximize variability within and between sets. Additionally, G5 was used in the SSS-test and the remaining four sets (G1 to G4) were used to create the stimuli (word streams, words and part-words) for the word-learning task.

Random syllable streams were created by randomly combining the syllables of a set, with no gap between them and with the sole constraint that the same syllable was not to be repeated consecutively. The total duration of the random syllable streams was of 60 seconds for the SSS-test (G5) and of 120 seconds for the MEG task (G1 to G5).

For the word-learning task, we created four tri-syllabic pseudowords (henceforth, words) per set (G1 to G4) by the unique combination of all of its comprising syllables. This resulted in four 'languages' (henceforth, languages; L1 to L4), each comprising four words. To improve learnability, we consulted the Irvine Phonotactic Online Dictionary (IPhOD version 2.0; www.IPhOD.com) for minimum word-average biphoneme and positional probabilities. For the exposure phase, the four words of a language were randomly combined to form a two-minute long stream for each language with no gaps between them, thus ensuring an equal number of non-consecutive repetitions per word. For the test phase, in addition to the words, we also created for each language all possible part-words by the combination of the final syllable of a word with the first two syllables of the remaining words (12 part-words per language). Written renderings and cross-ratings of all words and part-words were provided by five independent American English native speakers. The written forms with the highest convergence were selected for the visual presentation concurrent with the audio in the test of the word learning task.

Random syllable streams, word streams, words and part-words were all converted to .wav files for auditory playback using the MBROLA text-to-speech synthesizer with the American Male Voice diphone database (US2) at 16kHz[38]. All phonemes were equal in pitch (200Hz), pitch rise and fall (with the maximum at 50% of the phoneme), and duration, which was set to 111 ms to satisfy a presentation rate of 4.5 syllables per second throughout the entire study.

All auditory stimuli were presented binaurally at a mean 75 dB sound pressure, via tubephones (E-A-RTONE 3A 50Ω, Etymotic Research) attached to E-A-RLINK foam plugs inserted into the ear canal.

SSS-test

Participants (for the in-lab experiment in a sound isolation booth, seated in front of a PC with a microphone placed close to their mouth) completed three experimental steps:

i. *Volume adjustment*: Subjects listened to the train of random syllables played backwards while whispering *"tah"*, and increased the audio volume until they could not perceive their own voice.

ii. *Steady repetition example*: An audio with a continuous repetition of the syllable *"tah"* (recorded by a female speaker, manipulated to last 222ms and concatenated to produce a rate of 4.5 syllables per second) was delivered through the earplug tubephones for 10 seconds. Subsequently, participants were instructed to whisper *"tah"* at the same pace during 10 seconds. We primed the participants at the desired frequency, since previous research showed that synchronization to an external stimulation occurs when there is a close match between internal and external frequencies[39,40].

iii. *Syllable perception task*: Participants attended to the rhythmic syllable stream while steadily whispering *"tah"*. After the presentation, they had to indicate whether a given set of target syllables were presented. For each run, four target syllables were randomly selected from a pool of eight (half of them were part of the stream). Importantly, participants were not explicitly instructed to synchronize to the external audio. According to the instructions, the assignment was to recall correctly the syllables and the *"tah"* articulation was intended just to increase the difficulty of the task. By this, we encouraged attention to the audio while the goal of the task remained orthogonal to the synchronization (implicit synchronization). A subset of participants, randomly selected from the pool, completed an extra step at the end of the syllable perception task (N_{high} = 13, N_{low} = 12). During this step, they steadily whispered "*tah*" during one minute while listening to a white noise (no-rhythm condition).

After the last step, participants filled a questionnaire indicating age, handedness, gender, musical experience and spoken languages. Subsequently, they repeated steps (*ii*) and (*iii*). In this way, each participant completed two runs of the *(iii) Syllable perception task*, which we named blocks one and two, respectively.

A subgroup of participants, randomly selected from the original cohort, completed the whole experiment again one month after the first session (N = 34).

SSS-test, accelerated version—The protocol for the accelerated version was the same as in the regular SSS-test, but we modified the auditory stimulus as follows: we progressively increased the syllable rate from 4.3 to 4.7 syllables per second, using steps of 0.1 sylls/sec; each rate was kept constant for 60 syllables with the exception of the last one that remained constant until the end of the audio, which in this case was set to 70 seconds (see red trace in Fig. S2B). As in the regular version of the test, each participant completed 2 blocks. Once they finished, they indicated if they perceived an increment, decrement or no change in the rate of the presented syllables of each block. As for the stable rate version of the SSS-test, we computed the PLV between the envelopes of the produced utterances and the input signal.

Online version—An online version of the SSST-test (for both normal and accelerated versions of the paradigm) was developed using *oTree*, a Python-based framework for the development of controlled experiments on online platforms[41]. The online version mainly followed the same structure as the in-lab one (Volume adjustment, Steady repetition, Syllable perception), but with some changes: i) a microphone test phase was included before the volume adjustment phase; ii) several restrictions were placed to ensure that the participants actually did the task (e.g., during Steady repetition, a participant could not continue to the next page until she/he had heard the whole 10 second-long recording and had recorded himself/herself for another 10 seconds). All recordings were manually checked for errors (e.g., not using headphones, speaking loud instead of whispering, etc.). Instructions were exactly the same as those for the in-lab version.

Half of participants who completed the online version of the stable rate SSS-test, also completed a rhythm perception test, while the other half undertook a rhythm production assessment. The perception task is the one described by Huss and colleagues[42] Participants were presented with pairs of tunes (36 pairs, 18 equal and 18 different) and had to make a same/different judgment. In the different pairs, there is a mismatch in the duration of the accented note. For the production task, participants repeated the Steady repetition example step (see Methods), and they were instructed to keep whispering “tah” at the same pace for one minute while listening to a white noise. The difference between this task and the no-rhythm condition (see Methods) is that all participants were primed again, before the one minute of whispering under noise, and explicitly instructed to keep the tempo. These tasks were always performed after the SSS-test and were also programmed using oTree.

Tasks developed with *oTree* can be deployed into Amazon Mechanical Turk (AMT), a crowdsourcing platform which allows for the acquisition of large datasets in a rapid, affordable, anonymous, and easily replicable manner. Note that recent research has replicated a number of tasks from experimental psychology (e.g., Stroop, Flanker, subliminal priming and category learning among others) by using AMT[43]. The online SSSTtask was presented to AMT participants as an HTML webpage that ran in each participant's web browser. AMT participants were first presented with a summary of the task and then with an informed consent page. Upon acceptance, instructions for the task were presented.

Synchrony measurement—The degree of synchronization was measured by the phase locking value (PLV) between the envelope of the produced speech and the cochlear envelope of the rhythmic syllable stream. The envelope was estimated as the absolute value of the Hilbert transform of the signal. The spectrograms of the auditory stimuli were computed using the NSL (Neural Systems Laboratory) Auditory Model Matlab toolbox[44]. This toolbox filters the signal in a specific set of frequency bands, emulating the auditory filters applied by the cochlea (yielding what we call auditory channels), and computes their envelopes. The stimulus cochlear envelopes were calculated by adding the auditory channels between 180 and 7246 Hz. Envelopes were resampled at 100 Hz, filtered between 3.5 and 5.5 Hz and their phases were extracted by means of the Hilbert transform. The PLV was computed for windows of 5 seconds length with an overlap of 2 seconds. The results for all time windows were averaged within each stimulus presentation, providing one PLV per block.

We computed the within and across session PLV Spearman rank's correlation. For the within session blocks, the correlation was computed between the PLV of block 1 and the one of block 2. For the across session computation, the mean PLV across blocks for the first session was correlated with the one obtained for the session completed one month later.

Definition of high and low synchronizers—We applied a *k-mean* clustering algorithm[45] on the PLV values averaged across blocks, using a squared Euclidean distance metric and 2 clusters. For each cluster (the *lower* and the *higher* one) we fitted a normal distribution (mean -μ- and standard deviation -σ-). Next, we defined a *low* and a *high* threshold (T) as: $T_{low}=\mu_{lower}+\sigma_{lower}$ and $T_{high}=\mu_{higher}-\sigma_{higher}$. The groups of low and high synchronizers, which underwent the MEG and DW-MRI sessions, were randomly selected from the set of participants whose mean PLV across blocks was below T_{low} and above T_{high}, respectively.

Spectral analysis—We computed the Discrete Fourier Transform (DTF) for the envelope of the produced speech for each block without any windowing. We focused our analysis between 1 and 10 Hz. Thus, we just kept the power values within this frequency window and they were normalized to sum 1. Spectra across blocks belonging to the same condition (no-rhythm/rhythm) were averaged.

In order to assess differences between conditions (no-rhythm/rhythm), within each group (high/low synchronizers), non-parametric paired Wilcoxon Signed-Rank test were calculated for the power values at every frequency. Results are reported at a FDR-corrected $p < 0.05$ value.

Neurophysiological study

Task—Once in the MEG system, participants completed five runs of the *Syllable Perception Task*. Importantly, however, and in contrast to the SSS-test, participants passively listened to the syllable streams in silence (i.e. without whispering). Each run involved a different random syllable stream corresponding to syllable sets G1 to G5. Their order of presentation was counterbalanced across participants. Random syllable streams for the MEG experiment were of 120 seconds duration. Experimental runs were always preceded by 20 seconds of silence. Participants were instructed to listen attentively to the syllables and to indicate in the test phase after each stream was heard whether a given set of syllables had been presented. As in the SSS-test, each test phase consisted of eight trials. In each, a syllable from a pool of eight (four presented and four foils) was randomly selected and presented visually. Participants were to indicate their decision via a button press with their right hand (index finger: yes, the syllable was present; middle finger: no, the syllable was not present). The following trial started between 900 and 1100 ms after the response, for which there were no time constraints.

Data acquisition and processing—Neuromagnetic responses were recorded with a 1000 Hz sampling rate using a 157-channel whole-head axial gradiometer system (KIT, Kanazawa Institute of Technology, Japan) in a magnetically shielded room. Five electromagnetic coils were attached to the subject's head to monitor head position during MEG recordings. The coils were localized to the MEG sensors at the beginning of the experiment and before the last two blocks of the main experiment. The position of the coils with respect to three anatomical landmarks: the nasion, and left and right tragus were determined using 3D digitizer software (Source Signal Imaging, Inc.) and digitizing hardware (Polhemus, Inc.). This measurement allowed a coregistration of subjects' anatomical magnetic resonance image (MRI) with the MEG data. An online bandpass filter between 1 and 200 Hz and a notch filter at 60 Hz were applied to the MEG recordings.

Data processing and analyses were conducted using custom MATLAB code and the FieldTrip toolbox[46]. For each participant's dataset, noisy channels were visually rejected. Two procedures were applied to the continuous MEG recordings. First, a least squares projection was fitted to the data from the 2 minutes of *empty room* recorded at the end of each session. The corresponding component was removed from the recordings[47]. Second, the environmental magnetic field, measured with three reference sensors located away from the participant's head, was regressed out from the MEG signals using time-shifted PCA[48]. The MEG signals were then detrended and artifacts related to eyeblinks and heartbeats were removed using independent component analysis.

Source reconstruction—In order to reconstruct the brain activity generating the magnetic fields recorded by the MEG sensors, we used a linearly constrained minimum variance beamforming approach. Using the subject's anatomical MRI we first reconstructed the brain surface. Then, the brain volume was filled with a 1 cm grid, which was normalized to the MNI template (Montreal Neurological Institute brain) using SPM8 (www.fil.ion.ucl.ac.uk/spm). The lead fields were calculated for every grid point[49] and the spatial filters were computed using the covariance matrix between all sensor pairs for all the trials. Finally, the spatial filters were applied to the sensors' signals to reconstruct the time series for every source inside the brain (i.e. point of the grid).

Brain-to-stimulus synchronization—The degree of synchrony was measured by the PLV between brain activity and the *cochlear* envelope of the perceived train of syllables.

Sources' signals were resampled at 100 Hz, filtered between 3.5 and 5.5 Hz and their phases were extracted by means of the Hilbert transform. The PLV was computed for windows of 1 second duration with an overlap of 0.5 seconds. The results for all time windows were averaged across the total presentation of the stimuli, obtaining one PLV per source and per subject.

The PLVs were averaged for sources within the same region according to the Brainnetome Atlas[50] (38 mean PLV values were computed for the frontal ROI and 24 for the temporal one, see below). ROI selection was theoretically driven, based on our initial hypothesis related to the audio-motor nature of the behavioral task. In other words, since the segregation of participants relies on an audio-motor task, we focused our analyses on two broad regions, comprising the cortical areas related speech perception and production:

i. Bilateral frontal ROI: It is composed of 19 regions in each hemisphere, 38 in total. It comprises the middle, precentral and inferior frontal gyrus from both hemispheres. Specifically the selected Brainnetome Atlas regions are: dorsal BA 9/46, Inferior Frontal Junction, BA 46, ventral BA 9/46, ventro-lateral BA 8, ventro-lateral BA 6, lateral BA 10, dorsal BA 44, Inferior Frontal Sulcus, caudal BA 45, rostral BA 45, opercular BA 44, ventral BA 44, head and face region BA 4, caudal dorso-lateral BA 6, upper limb region BA 4, trunk region BA 4, tongue and larynx region BA 4, caudal ventro-lateral BA 6 (Fig. 2B).

ii. Bilateral temporal ROI: It consists of 12 regions in each hemisphere, 24 in total. This ROI's covers bilaterally the superior, middle and posterior temporal lobe (medial BA 38, BA 41/42, TE 1.0/1.2, caudal BA 22, lateral BA 38, rostral BA 22, caudal BA 21, rostral BA 21, dorsolateral BA 37, anterior Superior Temporal Sulcus -STS-, rostro-posterior STS and caudo-posterior STS; Fig. S6B).

Since the preference of right auditory areas for frequencies matching the syllable rate has been theoretically proposed[51,52] and experimentally reported[53–55], we also explored the degree of asymmetry in auditory entrainment to speech. Based on the previous literature[53,54], more restricted ROIs were chosen for this analysis. We defined early auditory regions as: BA 41/41, TE 1.0 and TE 1.2 (see Fig S7A). Next, we calculated brain-to-stimulus synchrony within right and left early auditory areas for each group and we calculated their neurophysiological asymmetry: $(PLV_{right} - PLV_{left})/0.5(PLV_{right} + PLV_{left})$.

In order to assess differences between groups (high/low synchronizers) non-parametric independent samples Mann-Whitney-Wilcoxon tests were calculated for all regions' PLVs within the corresponding ROI. Results are reported at a FDR-corrected $p < 0.05$ value within ROI. To explore auditory brain-to-stimulus synchrony between hemispheres, within groups, a non-parametric paired Wilcoxon Signed-Ranks test was computed.

For the frontal ROI, 8 regions from the 39 showed a significant difference between groups, FDR corrected for multiple comparisons. All significant regions were located in the left hemisphere, specifically left: dorsal BA 9/46, ventral BA 9/46, Inferior Frontal Junction, dorsal BA 44, ventral BA 44, opercular BA 44, Inferior Frontal Sulcus, caudal BA 45.

Anatomic connectivity study

Scanning Parameters and diffusion measures—DW-MRI data were acquired on a 3T scanner (Siemens Prisma 3T MRI scanner) using a 64-channel phased-array head coil, at the Center for Brain Imaging (New York University). Diffusion images were acquired with an EPI sequence optimized for DTI-MRI of white matter -WM- (81 axial slices, TR: 4150 ms, TE: 85.2 ms, flip angle: 90°, slice thickness: 1.5 mm, acquisition matrix: 150 × 152, voxel size: 1.5 × 1.5 × 1.5 mm^3). One run with ten interleaved non-diffusion weighted volumes and 128 diffusion weighted volumes (i.e., 128 directions; b-values of 1500 s/mm^2) was acquired. To allow a precise source reconstruction of the neurophysiological data, a high resolution T1 MPRAGE image was also acquired during this MRI session (TR = 2400 ms, TE = 2.24 ms, flip angle = 8°, voxel size = 0.80 × 0.80 × 0.80 mm^3, 256 sagittal slices, acquisition matrix = 320 × 300).

DTI-MRI analysis—Diffusion data processing started by correcting for eddy current distortions and head motion using FMRIB's Diffusion Toolbox (FDT), which is part of the FMRIB Software Library -FSL 5.0.1, www.fmrib.ox.ac.uk/fsl/[56]. Subsequently, the gradient matrix was rotated corresponding to the head movement, to provide a more accurate estimate of diffusion tensor orientations using the *fdt_rotate_bvecs* program included in FSL[57]. Brain extraction was performed using the Brain Extraction Tool[58], which is also part of the FSL distribution. The analysis continued with the reconstruction of the diffusion tensors using the linear least-squares algorithm included in Diffusion Toolkit 0.6.2.2[59]. Finally, Fractional Anisotropy (FA) and Radial Diffusivity (RD) maps for each participant were calculated using the eigenvalues extracted from the diffusion tensors.

Voxel based analyses of FA and RD maps were performed using Tract Based Spatial Statistics, TBSS[60]. FA maps from all participants were registered to the FMRIB58_FA template (MNI152 space and 1×1×1 mm^3) using the nonlinear registration tool[61]. These registered FA maps were first averaged to create a mean FA volume. Then a mean FA skeleton was derived, which represents the centers of all tracts common to all participants in the study. Each participant's aligned FA data were then projected onto this skeleton by searching for the highest FA value within a search space perpendicular to each voxel of the mean skeleton. This process was repeated for the RD maps by applying the transformations previously calculated with the FA maps. This resulted in individual FA and RD skeletons for each participant. In addition, given that laterality[62] —specially of WM paths connecting auditory and motor regions —is also related to cognitive function, laterality maps were also created. First a symmetric skeleton was created using the script *tbss_sym*. Then FA and RD data were projected to this symmetric skeleton, with left-hemispheric values being subtracted from right-hemispheric ones. Thus, laterality FA and RD maps were also obtained (note that these maps reflect right minus left values; for the sake of clarity, results are shown on the left hemisphere, Fig. 3). A total of 4 analyses were performed (FA and RD, FA and RD lateralization).

Finally, in order to assess white matter differences between high and low synchronizers, independent samples t-tests were calculated for the FA and RD skeletons and laterality maps. For a more theoretically driven analysis, we focused on the regions that are part of the dorsal and ventral pathways for language processing. In particular, we used a ROI approach to focus on WM pathways connecting auditory with frontal and motor regions -i.e., the arcuate fasciculus[21]. We also included in the analyses the ventral pathways connecting temporal, occipital and frontal areas as *control regions*. Our ROI was created using well-known probabilistic atlases of white matter pathways in MNI space[16,63]. We included any voxel within the skeleton, which had at least a 50% of being part of the long, anterior, and posterior segments of the arcuate fasciculus, the inferior-fronto occipital fasciculus, the inferior longitudinal fasciculus and the uncinate fasciculus. Results are reported at a FWE-corrected $p < 0.05$ value using threshold-free cluster enhancement[64] and a nonparametric permutation test with 5000 permutations[65]. Significant voxels within the skeleton were filled to make the presentation of results easier to follow. Significant clusters (results) were averaged and a mean value per participant, reflecting individual microstructural differences was obtained. These diffusion values were then correlated (using Spearman rank's correlation) to MEG derived measures of brain synchrony in both frontal and auditory regions (FDR-corrected for the 2 correlations computed; see MEG results).

Tractography analyses—Given the TBSS results showing that a cluster consistent with the left arcuate fasciculus differentiated between high and low synchronizers (Fig 3), confirmatory ad-hoc virtual dissections (i.e., deterministic tractography) were also computed to further locate the white matter pathway underlying the pattern of results. Specifically, for each participant we manually dissected the three segments (long, anterior and posterior) of the left arcuate fasciculus[24]. As a control, the left inferior-fronto-occipital fasciculus (IFOF) and the left inferior longitudinal fasciculus (ILF) were also dissected. The IFOF and ILF were selected based on: *(i)* the fact that their anatomy could partially overlap with the TBSS cluster; and *(ii)* research suggesting that these pathways are part of the ventral pathway for language processing[14,29,66].

Using the previously pre-processed diffusion data, whole-brain tractography was performed using Diffusion Toolkit 0.6.2.2[59] and the interpolated streamlines algorithm. Tractography was started only in voxels with an FA value greater than 0.2 and was stopped when the angle between two consecutive steps was larger than 35°. Manual dissection of the tracks was performed using Trackvis[59]. ROIs were defined using the T1 high resolution image and the fractional anisotropy (FA) and FA color-coded maps as a reference for individual anatomical landmarks. The three segments of the left arcuate were dissected using established guidelines with a two-sphere approach. For the two control ventral tracts, three spherical ROIs at the level of the anterior temporal lobe (temporal ROI), the posterior region located between the occipital and temporal lobe (occipital ROI) and the anterior floor of the external/extreme capsule (frontal ROI) were created. In order to define each of the ventral tracts of interest, we applied a two-ROI approach: the ILF was obtained by connecting the temporal and occipital ROIs, while the streamlines passing through the occipital lobe and frontal ROIs were considered as part of the IFOF. All these ROIs were applied according to a well-defined anatomical atlas[16]. The exclusion of single fiber structures that do not represent part of the dissected tract was achieved using subject specific no-ROIs. After the dissection was completed, the volume and the mean FA and RD value of each tract was extracted for further analysis. In order to take into account individual differences in head volume, volumes from all tracts were corrected by dividing their original value by each subject Total Intracranial Volume (TIV). TIV was calculated by submitting each participant's T1 high resolution image to the standard Freesurfer (http://surfer.nmr.mgh.harvard.edu/) pipeline. Using the extracted values (FA, RD and volume), we computed between group comparisons for the arcuate as a whole (sum of anterior, long and posterior segments), its three segments separated and the IFOF and ILF as control tracts. Thus we computed a total of 18 comparisons. We used a $p < 0.05$ FDR correction threshold to correct for these multiple calculations.

Phonological word-form learning task

The task consisted of a *Volume adjustment* step as above and two runs with the following experimental steps each:

- *Exposure phase*: participants were exposed to a two-minute long speech stream of words corresponding to one of the created languages (L1-L4, see Table S1) and were asked to remain silent (i.e. no whispering) during the auditory presentation.
- *Test phase:* Each word stream was immediately followed by a test phase. Test trials consisted of a two alternative forced choice between a word and a part-word, both randomly selected from the pool corresponding to their language. Each word and part-word appeared only twice, each time paired with a different item. The total number of test trials was thus 8. Test items were presented both in their auditory and written forms and were assigned a number (1 or 2) according to their auditory presentation and left-right presentation on the screen. Participants were asked to make their choice by pressing the corresponding number.

The language presentation order was counterbalanced across participants. All participants thus completed two runs, each testing a different language. The proportion of correct responses in the two runs was averaged before proceeding with group analyses.

The group of participants (N=100) who completed the online and accelerated version of the SSS-tests, also completed an online version of this word-form learning task with the above specifications. The online word learning task was also created using oTree.

In order to assess learning differences between groups (high/low synchronizers) a non-parametric independent samples Mann-Whitney-Wilcoxon test was calculated for the averaged proportion of correct responses.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. Correspondence and request for materials should be addressed to MFA (fassaneo@gmail.com)

Code availability

All computer code used for this study is available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

We thank Jess Rowland and Jean-Remi King for comments and advice. This work was supported by NIH 2R01DC05660 (DP) and FP7 Ideas: European Research Council ERC-StG-313841 (RDB).

References

1. Condon WS & Sander LW Neonate movement is synchronized with adult speech: interactional participation and language acquisition. Science 183, 99–101 (1974). [PubMed: 4808791]
2. Repp BH Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin and Review 12, 969–992 (2005). [PubMed: 16615317]
3. Woodruff Carr K, White-Schwoch T, Tierney AT, Strait DL & Kraus N Beat synchronization predicts neural speech encoding and reading readiness in preschoolers. Proc. Natl. Acad. Sci 111, 14559–14564 (2014). [PubMed: 25246562]
4. Patel AD The Evolutionary Biology of Musical Rhythm: Was Darwin Wrong? PLoS Biol 12, (2014).
5. Janata P & Grafton ST Swinging in the brain: Shared neural substrates for behaviors related to sequencing and music. Nature Neuroscience 6, 682–687 (2003). [PubMed: 12830159]
6. Merchant H, Grahn J, Trainor L, Rohrmeier M & Fitch WT Finding the beat: a neural perspective across humans and non-human primates. Philos. Trans. R. Soc. B Biol. Sci 370, 20140093–20140093 (2015).
7. Assaneo MF & Poeppel D The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. Sci. Adv 4, eaao3842 (2018).
8. Park H, Ince RAA, Schyns PG, 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). [PubMed: 26028433]
9. Steele CJ, Bailey JA, Zatorre RJ & Penhune VB Early Musical Training and White-Matter Plasticity in the Corpus Callosum: Evidence for a Sensitive Period. J. Neurosci 33, 1282–1290 (2013). [PubMed: 23325263]
10. Varnet L, Ortiz-Barajas MC, Erra RG, Gervain J & Lorenzi C A cross-linguistic study of speech modulation spectra. J. Acoust. Soc. Am 142, 1976–1989 (2017). [PubMed: 29092595]
11. Ding N, Patel A, Chen L, Butler H, Luo C, & Poeppel D Temporal Modulations in Speech and Music. Neurosci. Biobehav. Rev (2017).
12. Long MA et al. Functional Segregation of Cortical Regions Underlying Speech Timing and Articulation. Neuron 89, 1187–1193 (2016). [PubMed: 26924439]
13. Magrassi L, Aromataris G, Cabrini A, Annovazzi-Lodi V & Moro A Sound representation in higher language areas during language generation. Proc. Natl. Acad. Sci 112, 1868–1873 (2015). [PubMed: 25624479]
14. Ripollés P et al. Strength of temporal white matter pathways predicts semantic learning. J. Neurosci 1720–17 (2017). doi:10.1523/JNEUROSCI.1720-17.2017
15. Lopez-Barroso D et al. Word learning is mediated by the left arcuate fasciculus. Proc. Natl. Acad. Sci 110, 13168–13173 (2013). [PubMed: 23884655]
16. Thiebaut de Schotten M et al. Atlasing location, asymmetry and inter-subject variability of white matter tracts in the human brain with MR diffusion tractography. Neuroimage 54, 49–59 (2011). [PubMed: 20682348]

17. Lopez-Barroso D et al. Language learning under working memory constraints correlates with microstructural differences in the ventral language pathway. Cereb. Cortex 21, 2742–2750 (2011). [PubMed: 21527790]

18. Saffran JR, Aslin RN & Newport EL Statistical learning by 8-month-old infants. Science (80-.) 274, 1926–1928 (1996).

19. Morillon B & Baillet S Motor origin of temporal predictions in auditory attention. Proc. Natl. Acad. Sci 114, E8913–E8921 (2017). [PubMed: 28973923]

20. Cummins F Rhythm as entrainment: The case of synchronous speech. J. Phon 37, 16–28 (2009).

21. Hickok G & Poeppel D The cortical organization of speech processing. Nature Reviews Neuroscience 8, 393–402 (2007). [PubMed: 17431404]

22. Rauschecker JP & Scott SK Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. Nature Neuroscience 12, 718–724 (2009). [PubMed: 19471271]

23. Pulvermüller F & Fadiga L Active perception: Sensorimotor circuits as a cortical basis for language. Nature Reviews Neuroscience 11, 351–360 (2010). [PubMed: 20383203]

24. Catani M, Jones DK & Ffytche DH Perisylvian language networks of the human brain. Ann. Neurol 57, 8–16 (2005). [PubMed: 15597383]

25. Zatorre RJ, Fields RD & Johansen-Berg H Plasticity in gray and white: Neuroimaging changes in brain structure during learning. Nature Neuroscience 15, 528–536 (2012). [PubMed: 22426254]

26. Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA & Poeppel D Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron 93, 480–490 (2017). [PubMed: 28182904]

27. Hage SR & Nieder A Dual Neural Network Model for the Evolution of Speech and Language. Trends in Neurosciences 39, 813–829 (2016). [PubMed: 27884462]

28. Guenther FH Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol. Rev 102, 594–621 (1995). [PubMed: 7624456]

29. Turken AU & Dronkers NF The Neural Architecture of the Language Comprehension Network: Converging Evidence from Lesion and Connectivity Analyses. Front. Syst. Neurosci 5, (2011).

30. Fries P Rhythm for Cognition: Communication Through Coherence. Neuron 88, 220–235 (2015). [PubMed: 26447583]

31. Bengtsson SL et al. Extensive piano practicing has regionally specific effects on white matter development. Nat. Neurosci 8, 1148–1150 (2005). [PubMed: 16116456]

Methods-only references

32. Pernet CR, Wilcox R & Rousselet GA Robust correlation analyses: False positive and power validation using a new open source matlab toolbox. Front. Psychol 3, (2013).

33. Rousseeuw PJ & Van Driessen K A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223 (1999).

34. Verboven S & Hubert M LIBRA: A MATLAB library for robust analysis. Chemom. Intell. Lab. Syst 75, 127–136 (2005).

35. Rousselet GA & Pernet CR Improving standards in brain-behavior correlation analyses. Front. Hum. Neurosci 6, (2012).

36. Cureton EE Rank-biserial correlation. Psychometrika 21, 287–290 (1956).

37. Kerby DS The Simple Difference Formula: An Approach to Teaching Nonparametric Correlation. Compr. Psychol 3, 11.IT.3.1 (2014).

38. Dutoit T & Pagel V Le projet MBROLA: Vers un ensemble de synthétiseurs vocaux disponibles gratuitement pour utilisation non-commerciale. in XXIèmes Journées d'Etude sur la Parole 441–444 (1996).

39. Ermentrout GB & Rinzel J Beyond a pacemaker's entrainment limit: phase walk-through. Am. J. Physiol 246, R102–6 (1984). [PubMed: 6696096]

40. Strogatz SH Nonlinear Dynamics and Chaos. Library 48, 498 (1994).

41. Chen DL, Schonger M & Wickens C oTree-An open-source platform for laboratory, online, and field experiments. J. Behav. Exp. Financ (2016). doi:10.1016/j.jbef.2015.12.001

42. Huss M, Verney JP, Fosker T, Mead N & Goswami U Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. Cortex 47, 674–689 (2011). [PubMed: 20843509]

43. Crump MJC, McDonnell JV & Gureckis TM Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. PLoS One (2013). doi:10.1371/journal.pone.0057410

44. Chi T, S. S NSL Matlab Toolbox Maryl. Neural Syst. Lab., Univ. Maryland.

45. Arthur D & Vassilvitskii S k-means++: the advantages of careful seeding. in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms 1027–1025 (2007). doi: 10.1145/1283383.1283494

46. Oostenveld R, Fries P, Maris E & Schoffelen JM FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput. Intell. Neurosci 2011, (2011).

47. Adachi Y, Shimogawara M, Higuchi M, Haruta Y & Ochiai M Reduction of non-periodic environmental magnetic noise in MEG measurement by Continuously Adjusted Least squares Method. in IEEE Transactions on Applied Superconductivity 11, 669–672 (2001).

48. de Cheveigné A & Simon JZ Denoising based on time-shift PCA. J. Neurosci. Methods 165, 297–305 (2007). [PubMed: 17624443]

49. Nolte G The magnetic lead field theorem in the quasi-static approximation and its use for magnetoenchephalography forward calculation in realistic volume conductors. Phys. Med. Biol 48, 3637–3652 (2003). [PubMed: 14680264]

50. Fan L et al. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cereb. Cortex 26, 3508–3526 (2016). [PubMed: 27230218]

51. Poeppel D The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. in Speech Communication 41, 245–255 (2003).

52. Zatorre RJ & Belin P Spectral and temporal processing in human auditory cortex. Cereb. Cortex 11, 946–53 (2001). [PubMed: 11549617]

53. Boemio A, Fromm S, Braun A & Poeppel D Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat. Neurosci 8, 389–395 (2005). [PubMed: 15723061]

54. Giraud AL et al. Endogenous Cortical Rhythms Determine Cerebral Specialization for Speech Perception and Production. Neuron 56, 1127–1134 (2007). [PubMed: 18093532]

55. Telkemeyer S et al. Sensitivity of Newborn Auditory Cortex to the Temporal Structure of Sounds. J. Neurosci (2009). doi:10.1523/JNEUROSCI.1246-09.2009

56. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW & Smith SM FSL. *Neuroimage* 62, 782–790 (2012). [PubMed: 21979382]

57. Leemans A & Jones DK The B-matrix must be rotated when correcting for subject motion in DTI data. Magn. Reson. Med 61, 1336–1349 (2009). [PubMed: 19319973]

58. Smith SM Fast robust automated brain extraction. Hum. Brain Mapp 17, 143–155 (2002). [PubMed: 12391568]

59. Wang R, Benner T, Sorensen AG & Wedeen VJ Diffusion Toolkit: A Software Package for Diffusion Imaging Data Processing and Tractography. Proc. Intl. Soc. Mag. Reson. Med 15, 3720 (2007).

60. Smith SM, Jenkinson M & Johansen-Berg H Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487–1505 (2006). [PubMed: 16624579]

61. Andersson JLR, Jenkinson M & Smith S Non-linear registration, aka spatial normalisation. FMRIB Technial Report TR07JA2 Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK (2007).

62. Catani M et al. Symmetries in human brain language pathways correlate with verbal recall. Proc. Natl. Acad. Sci 104, 17163–17168 (2007). [PubMed: 17939998]

63. Rojkova K et al. Atlasing the frontal lobe connections and their variability due to age and education: a spherical deconvolution tractography study. Brain Struct. Funct 221, 1751–1766 (2016). [PubMed: 25682261]

64. Smith SM & Nichols TE Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44, 83–98 (2009). [PubMed: 18501637]

65. Nichols TE & Holmes AP Nonparametric permutation tests for functional neuroimaging: A primer with examples. Hum. Brain Mapp 15, 1–25 (2002). [PubMed: 11747097]

66. Duffau H et al. New insights into the anatomo-functional connectivity of the semantic system: A study using cortico-subcortical electrostimulations. Brain (2005). doi:10.1093/brain/awh423

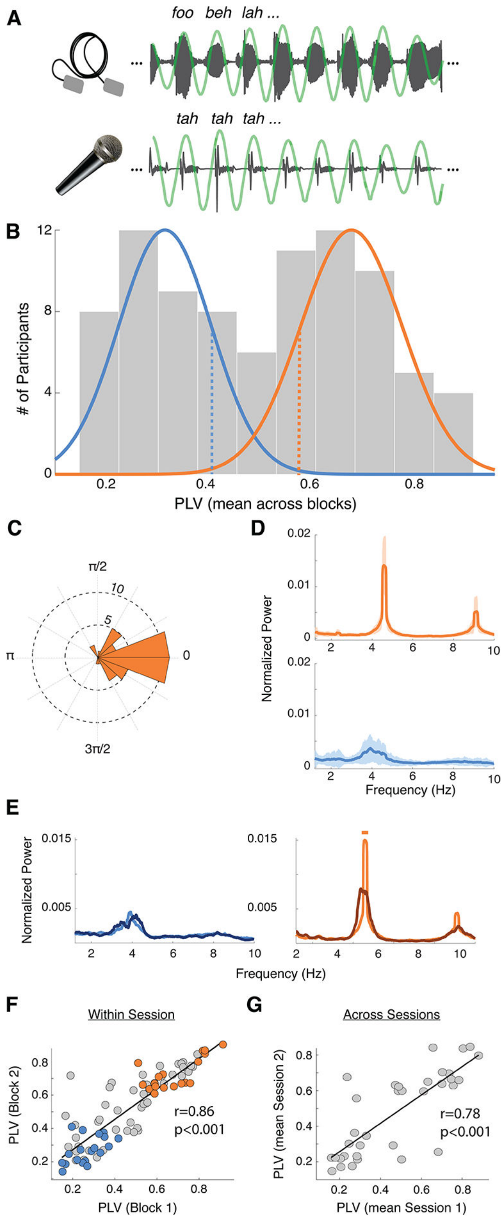

Fig. 1. Spontaneous speech synchronization reveals a bimodal distribution.

(**A**) SSS-test: example of the perceived (upper panel) and produced (lower panel) signals. Produced signals were independently recorded for each participant (N=84). Green line: the envelope bandpass filtered between 3.5–5.5 Hz. To eliminate auditory interference induced by listeners' own speech output, participants wore foam earplugs and whispered softly. (**B**) PLV histogram (average across blocks). Colored lines: normal distributions fitted to each of the two clusters obtained by a *kmeans* algorithm (the number of participants in each cluster is: N_{high} = 43, N_{low} = 41), low (blue)/high (orange) groups. Participants subsequently completing neurophysiology and neuroimaging sessions were randomly selected from below/above one sigma from the mean (blue/orange dashed lines). (**C**) Phase histogram for the lag between perceived and produced syllables. Histogram computed just for the high group. Low participants are not synchronized, thus it is not possible to define a phase lag.

(**D**) Average spectra of the utterances' envelopes (N_{high} = 43, N_{low} = 41). Shadowed regions: SD. (**E**) Average spectra for a subgroup of participants (N_{high} = 13, N_{low} = 12). Dark/light lines correspond to no-rhythm/rhythm conditions. Straight lines on top: significant difference between conditions (Wilcoxon signed rank test, two-sided $p < 0.05$, FDR-corrected). (**F**) PLV scatter plot of the correlation between first and second blocks (Spearman correlation coefficient $r = 0.86$, $p < 0.001$). Dots: individual subjects (N=84). Colored dots: participants selected to complete subsequent neurophysiology and neuroimaging sessions (N_{high} = 18, N_{low} = 19). (**G**) Scatterplot of the correlation between the mean PLV in the first and second sessions (one month apart; Spearman correlation coefficient $r = 0.78$, $p < 0.001$). Orange/blue correspond to high/low synchronizers, respectively, in all panels.

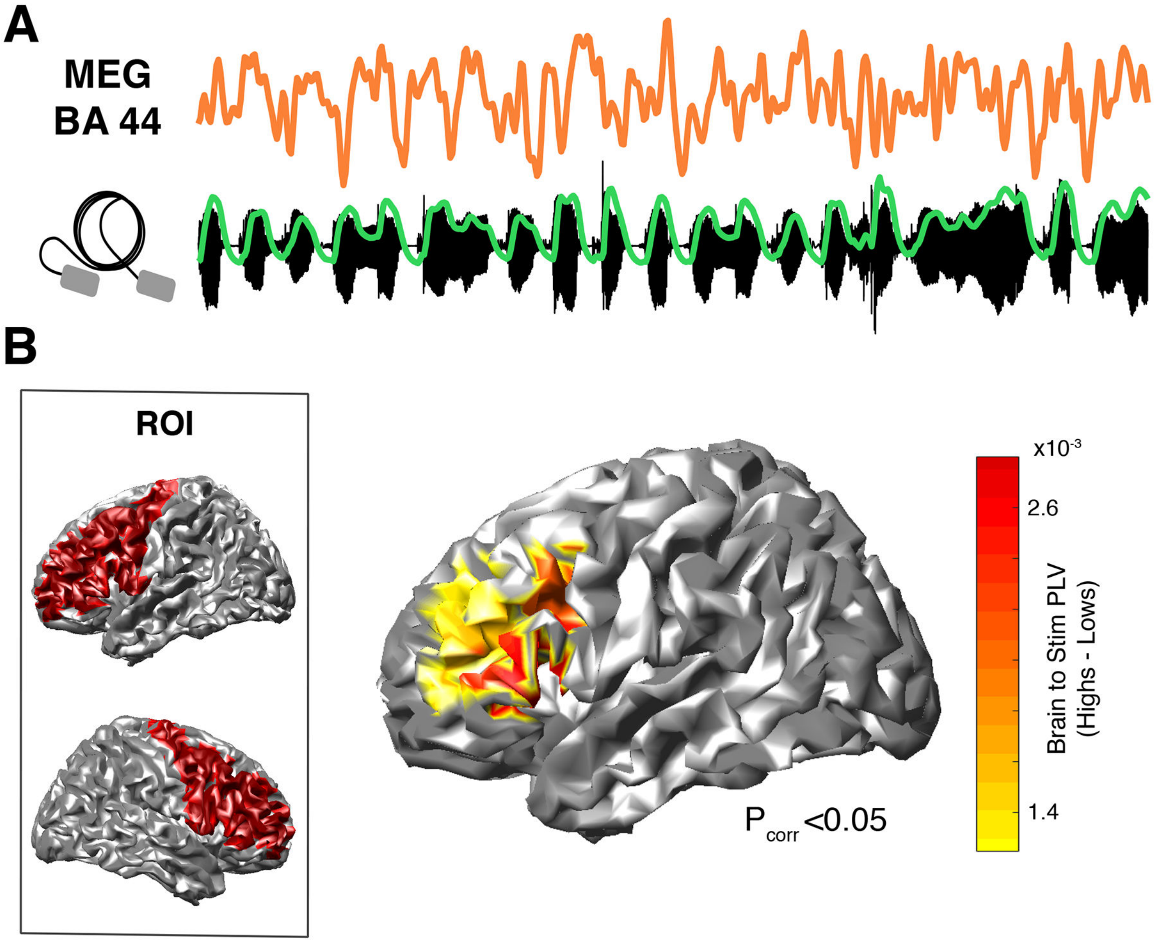

Fig. 2. Neural distinction between groups: neurophysiological data.

(A) Activity from a high synchronizer generated in BA44 (upper panel) during passive listening to the stimulus (lower panel; in green, stimulus envelope). Similar signals were obtained for the others (N=17) high synchronizers. **(B)** Brain-to-stimulus synchronization. Left panel: ROI comprising bilateral pre-central, middle frontal and inferior frontal gyri. Right panel: Brain surface map showing PLV *differences* between groups (N_{high} = 18, N_{low} = 19; Mann-Whitney-Wilcoxon test, two-sided p < 0.05, FDR-corrected).

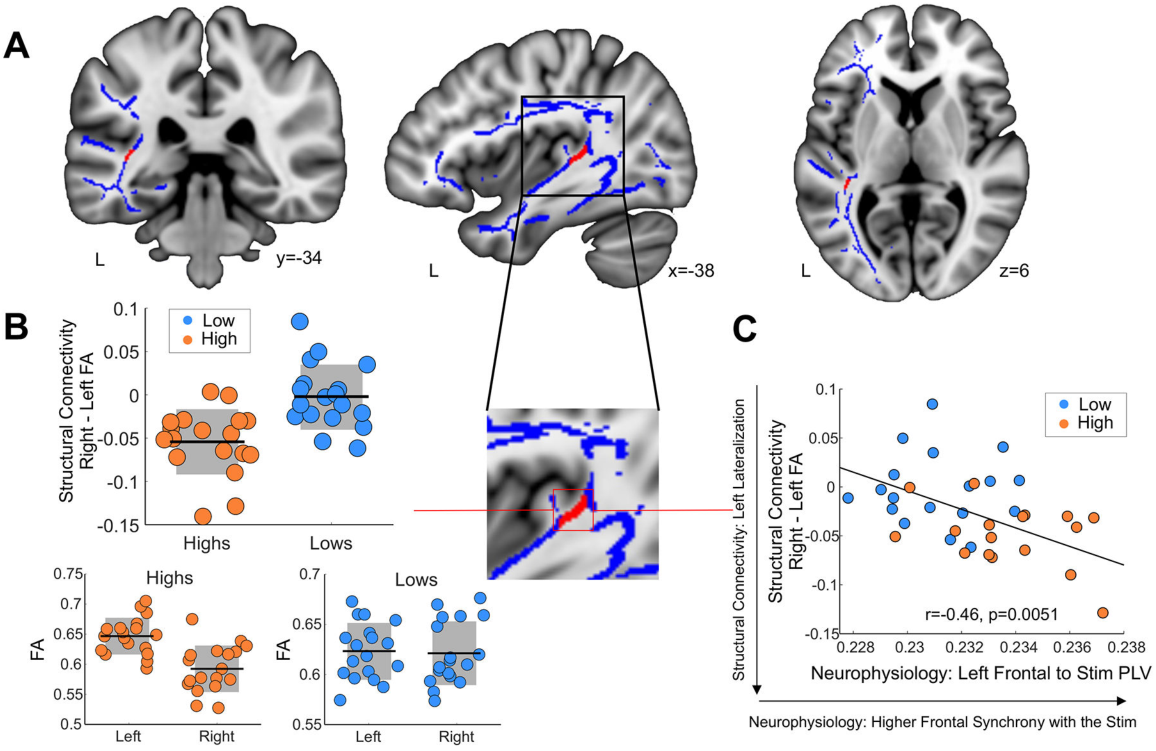

Fig. 3. Structural distinction between groups: anatomic connectivity data.

(A) Laterality maps (using tract based spatial statistics, TBSS) of Fractional Anisotropy (FA), right minus left values. In red, white matter pathways differentiating between groups (N_{high} = 18, N_{low} = 18, FWE-corrected, two-sided $p < 0.05$ using threshold free cluster enhancement) over the mean group skeleton (blue). Neurological convention is used, with MNI coordinates at the bottom of each slice. (**B)** To facilitate the visualization of the pattern of results, box-plots (N_{high}=18, N_{low}=18) with the mean (center line) and SD (grey areas) FA value of the significant cluster for each participant are shown for the laterality (right-left, top) and for each group and hemisphere separately (bottom). (**C**) Scatter plots (N=36) display the correlation between mean FA laterality values (negative values imply a leftward structural lateralization) and the synchrony of the left inferior/middle frontal gyri with the speech syllable rate **(**Spearman r = −0.46, p=0.0051; Skipped Spearman r = −0.44, t = −2.90, CI = −0.14, −0.68**)**. Orange/light blue, high/low synchronizers. Dots: individual participants. Black lines: mean across participants. Shadowed region: SD.

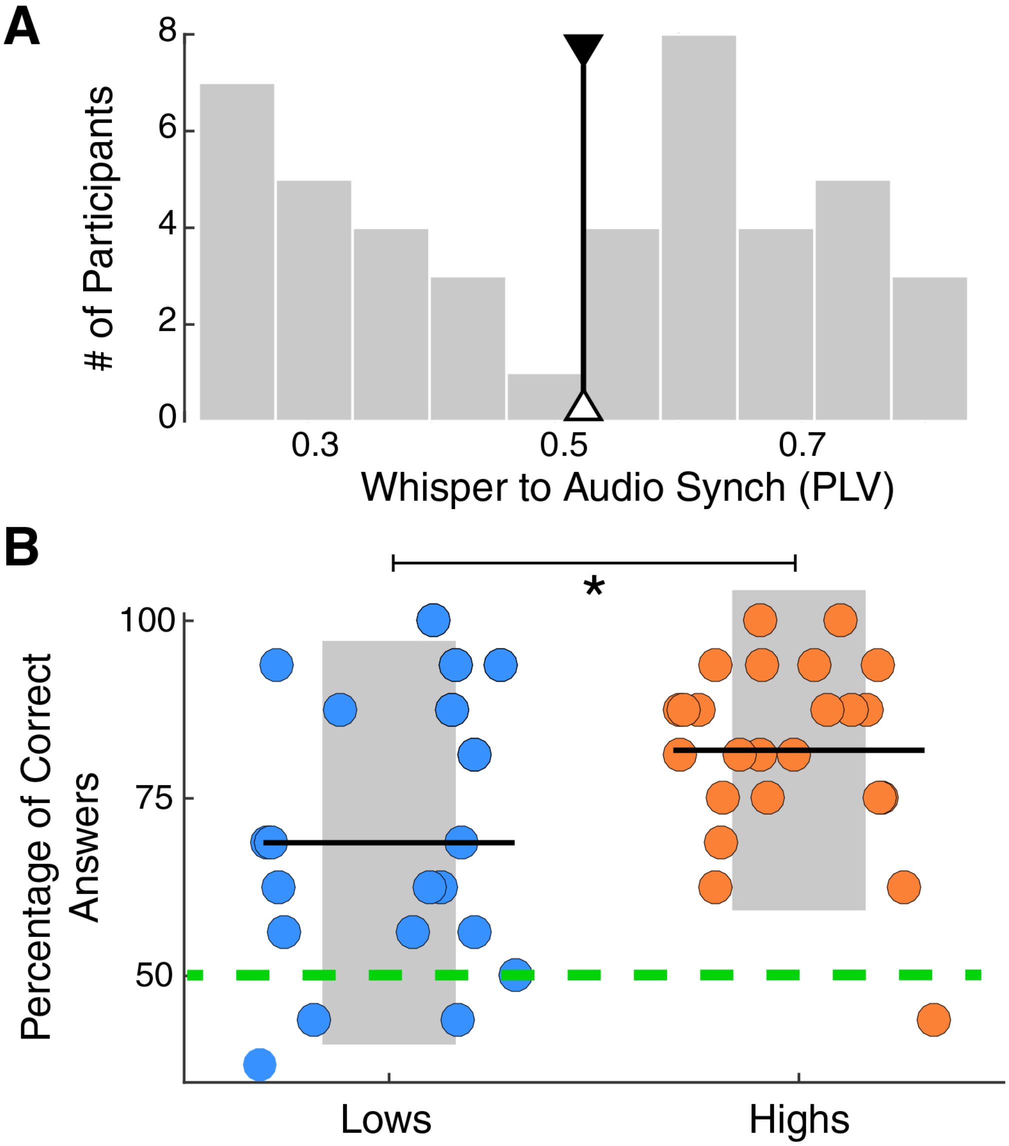

Fig. 4. Spontaneous Speech Synchronization test predicts word learning.

(**A**) SSS-test outcome. Histogram of the PLVs between the envelope of the perceived and produced speech signals, bandpass filtered at 3.5–5.5 Hz. The median of the first cohort's distribution is displayed (black line; individuals above/below this line are labeled as high/low). (**B**) Percent correct answers for the statistical word-learning task (N_{high} =24, N_{low}=20; Mann-Whitney-Wilcoxon test, two-sided p=0.024). Orange/light blue correspond to high/low synchronizers. Dots: individual participants. Black lines: mean across participants. Asterisk: p<0.05. Shadowed region: SD. Green dashed line: chance level in a two alternative forced-choice post-learning task.