# Population-level differences in the neural substrates supporting Statistical Learning

M. Florencia Assaneo[1,*,†], Joan Orpella[1,†], Pablo Ripollés[1], Laura Noejovich[1], Diana López-Barroso[2,3], Ruth de Diego-Balaguer[4,5,6,7,‡], David Poeppel[1,8,‡]

[1]Department of Psychology, New York University, New York, NY, USA

[2]Cognitive Neurology and Aphasia Unit, Centro de Investigaciones Médico-Sanitarias and Instituto de Investigación Biomédica de Málaga, University of Málaga, Málaga, Spain.

[3]Department of Psychobiology and Methodology of Behavioural Sciences, Faculty of Psychology and Speech Therapy, University of Málaga, Málaga, Spain.

[4]Cognition and Brain Plasticity Unit, IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain

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

[6]Institute of Neuroscience, University of Barcelona, Barcelona, Spain

[7]ICREA, Barcelona, Spain

[8]Neuroscience Department, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

* Corresponding author

† These authors contributed equally

‡ These authors jointly supervised this work

The ability to extract regularities from the environment is arguably an adaptive characteristic of intelligent systems. In the context of speech, statistical word-learning is thought to be an important mechanism for language acquisition. By taking into account individual differences in speech auditory-motor synchronization, an independent component analysis of fMRI data reveals that the neural substrates of this cognitive ability are not shared across individuals. While a network of auditory and superior pre/motor regions is universally activated to produce learning, a fronto-parietal network is instead additionally and selectively engaged by some individuals, boosting their performance. Interfering with the use of this network via articulatory suppression (producing irrelevant speech during learning) normalizes performance across the entire sample. Crucially, the engagement of this network predicts speech auditory-motor synchrony, directly relating this cognitive skill with language abilities.

## INTRODUCTION

Statistical learning (SL) is the capacity to use distributional information present in the environment to extract meaningful regularities. SL has been demonstrated across age groups from birth[1,2], sensory modalities (e.g., audition[3,4], vision[5] and touch[6]), domains[5] (temporal, spatial), and even species[7,8]. Specifically, statistical *word form* learning (SWFL) is widely used to examine the early stages of language acquisition by assessing the ability to derive phonological units, such as words, from continuous speech[3,9]. While several studies have investigated this ubiquitous cognitive phenomenon, its precise underlying neural substrates remain unclear.

While there have been several attempts to pinpoint the neural substrates of SWFL, the existing literature reveals non-overlapping results. Some studies found that SWFL performance correlates with the activity of superior temporal gyrus (STG) and dorsal pre/motor regions[10–12]. Other works have instead implicated the left inferior frontal gyrus[13] (IFG) and its interaction with superior temporal areas[14,15]. Here, we hypothesize that this variety of results is a consequence of individual differences in the neural resources allocated for SWFL.

Indeed, in a previous study, we revealed that individuals separated by their spontaneous speech auditory-motor synchronization abilities also differ in SWFL performance[16]. Specifically, we introduced a behavioral task (the Spontaneous Speech Synchronization test, SSS-test) that robustly classifies participants into high and low synchronizers. The high/low distinction reflected underlying neurophysiological and neuroanatomical individual differences and, critically, was predictive of SWFL performance (Fig. 1A). At the brain level, high synchronizers showed a greater brain-to-stimulus synchrony in the left IFG during passive speech listening and more volume in the white matter pathways underlying the dorsal language stream (i.e., the arcuate fasciculus)[17]. In line with this, we hypothesize that the recruitment of a network connected by this dorsal language pathway, including the IFG, gives high synchronizers the advantage in SWFL.

To test this hypothesis, we leveraged on a classic behavioral paradigm designed to interfere with the availability of the dorsal language stream for SWFL[18]. Such a paradigm involves the contrast between passive listening (PL) and articulatory suppression (AS) conditions (Fig. 1B). AS requires participants to repeat a nonsense syllable during learning, which decreases performance[9]. Although affine to a dual task paradigm, the executive load imposed by the

repeated syllable articulation is assumed to be minimal given the highly automatized nature of the articulation subtask. In the current work, we investigate the behavioral consequences of AS for high and low synchronizers and the different brain networks engaged by both synchrony groups during SWFL under PL and AS conditions.

To investigate the neural substrates of SWFL, we used Independent Component Analysis (ICA) of fMRI data. ICA is a data-driven approach well-suited to identify spatially independent and temporally coherent brain networks that support specific cognitive processes[19]. Previous work using this approach has related SWFL during PL to a network comprising auditory and superior pre/motor areas[12]. This work, however, did not take into account the high/low synchronizer distinction. We hypothesize that, if high synchronizers show better learning due to a greater reliance on the dorsal language stream, they should be more affected by AS and engage this stream differently from low synchronizers.

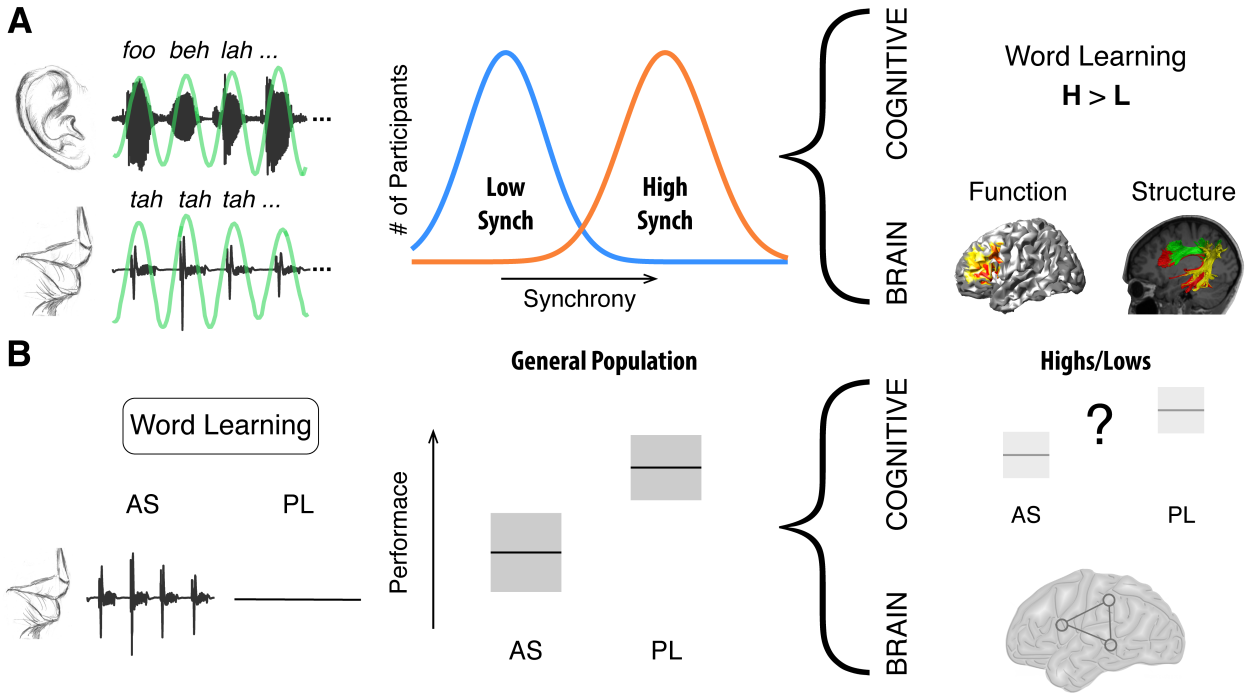


**Figure 1. Previous work motivating the current hypothesis. (A)** During the SSS-test, participants listen to an isochronous stream of random syllables (4.5 syll/sec) while concurrently whispering the syllable "*tah*". *Left inset:* example of the perceived (upper panel) and produced (lower panel) signals. Green line, bandpass filtered envelope used to compute input-output synchrony. *Middle panel:* Synchrony between perceived and produced syllables yields a bimodal distribution, allowing the classification of participants into low (blue) and high (orange) synchronizers. *Right panel:* High synchronizers

outperformed lows in a statistical word-learning task. They also showed enhanced brain-to-speech synchronization over left frontal regions and a greater volume in the white-matter pathways connecting temporal and frontal areas[16]. (**B)** The word-learning task consists in a learning phase where four tri-syllabic pseudo-words are presented in a continuous stream. Learning is assessed post exposure. *Left inset:* Participants are instructed to repeat a nonsense syllable (articulatory suppression condition) or listen (passive listening condition) during the entire learning phase. *Middle inset:* performance decreases due to articulatory suppression[9]. *Right inset:* differences between high and low synchronizers are hypothesized at the cognitive and brain levels.

## RESULTS

*Behavioral results: Articulatory Suppression modulates high but not low synchronizers' SWFL performance*

An initial cohort (N=56, 34 females; mean age, 22; age range, 18 to 37) underwent behavioral testing. Participants completed four blocks of statistical word-learning in two different experimental conditions, PL and AS, followed by the SSS-test (Methods and Fig. 1). In both tasks, the auditory stimuli were presented at a rate of 4.5 syllables per second, corresponding to the mean syllable rate across languages[20–22] and the natural frequency of speech-motor regions[23]. The outcome of the SSS-test showed the expected[16] bimodal distribution, allowing the classification of participants into high and low synchronizers (Fig. 2A). Moreover, the synchrony between perceived and produced syllables in the SSS-test was highly correlated with that in the AS blocks (Fig. 2B; N=56, Spearman correlation coefficient $r=0.73$, $p<0.005$). This demonstrates that speech-to-speech synchrony is not only reliable across *time*, as we previously demonstrated[16], but also across *tasks*, confirming that auditory-motor synchrony is a stable feature of each individual.

A first comparison of the learning performance between conditions over the entire sample (pooling high and low synchronizers together) showed a significant decrement in AS relative to PL (Fig. 2C; N=56, Wilcoxon signed-rank, two-sided $p=0.034$). This result thus replicates previously reported AS effects on SWFL[9]. When the sample was split into high and low synchronizers, the AS effect was observed only in the population of high synchronizers (Fig. 2D, orange color; $N_{high}=24$, Wilcoxon signed-rank, two-sided $p=0.026$). Strikingly, this effect *was not present* in low synchronizers: the performance of low synchronizers was not modulated by the action of speaking during the learning phase (Fig. 2D, blue color; $N_{low}=32$, Wilcoxon signed-

rank, two-sided p=0.37, $BF_{01}$=3.2). Additionally, and in line with our previous data[16], high synchronizers outperformed lows in the PL condition (Mann-Whitney-Wilcoxon test, two-sided p =0.039), but there was no difference in performance between groups in the AS condition (Mann-Whitney-Wilcoxon test, two-sided p=0.46, $BF_{01}$=3.1). Importantly, for all groups and conditions, learning did remain significantly above chance (Mann-Whitney-Wilcoxon test, two-sided $p_{high/AS}<0.005$, $p_{high/PL}<0.005$, $p_{low/AS}<0.005$, $p_{low/PL}<0.005$).

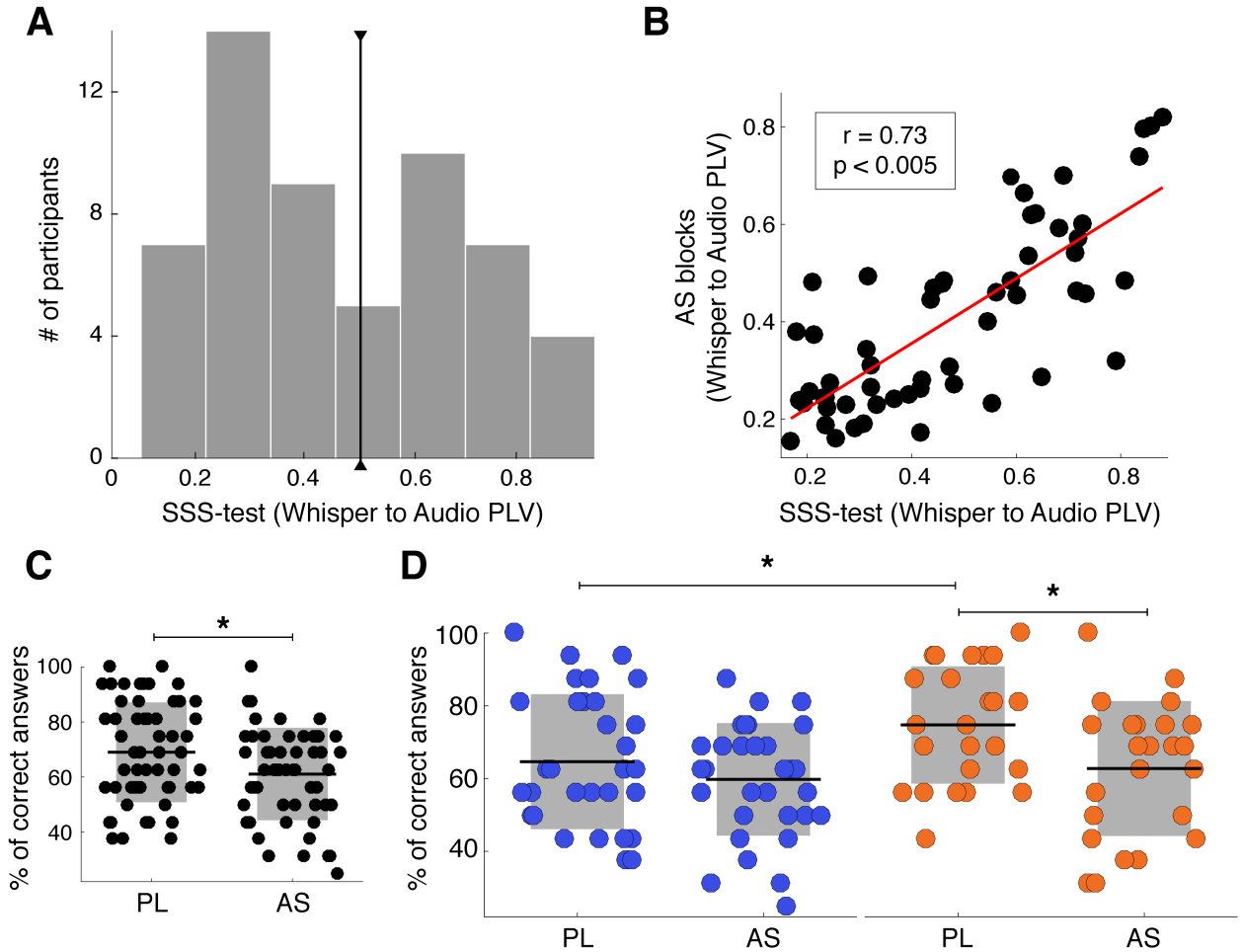


**Figure 2. Articulatory Suppression does not modulate low synchronizers' performance. (A)** SSS-test outcome. Histogram of the phase locking values (PLV; measure of speech-to-speech synchrony, see Methods) between the envelope of the perceived and produced speech signals, bandpass filtered at 3.5-5.5 Hz. Black line, critical value separating high and low synchronizers (N=56, see Methods). (**B**) Participants' PLV during AS as a function of the PLV from the SSS-test. Red line represents the correlation of the data. (**C**) Percentage of correct responses during PL and AS on the entire sample. (**D**) Percentage of correct responses during PL and AS for the low (blue) and the high (orange) synchronizers. Dots represent individual participants. *p<0.05. Mann-Whitney-Wilcoxon for between-group and

Wilcoxon Signed-Rank for within-group comparisons. Black lines: mean across participants. Shadowed regions: SD. AS: articulatory suppression; PL: passive listening.

*Neuroimaging results (I): High synchronizers activate an additional brain network during statistical word learning*

Having established the existence of behavioral differences between high and low synchronizers, we next acquired fMRI data from new participants (N=41) while they performed the same behavioral paradigm (see Method). The SWFL paradigm was optimized for fMRI testing. Specifically, we included a Resting and a Speech motor block as control conditions. During the Speech motor block, participants were required to repeatedly whisper the syllable *"tah"* with no auditory input. Again, the behavioral performance into the scanner replicated the learning pattern obtained with the first sample (Fig. S1). This replication under notably adverse learning conditions (during fMRI scanning), emphasizes the robustness of the behavioral results.

From the obtained images, we extracted 20 networks using the Group ICA of fMRI Toolbox[19] (GIFT; see Methods), 8 of which reflected artifacts related to, for example, movement, the presence of ventricles or blood vessels. From the 12 remaining networks, 5 were significantly recruited during statistical word learning in the PL and/or AS condition (Fig. S2 and Fig. S3). Critically, a fronto-parietal network including bilateral inferior and middle frontal gyri, inferior parietal cortex and the supplementary motor area distinguished between high and low synchronizers during the PL condition (Fig. 3A; $N_{high} = 18$, $N_{low} = 20$, Mann-Whitney-Wilcoxon test, two-sided p=0.038, FDR-corrected). Moreover, while the activity of this network during PL was significant for high synchronizers, it was not for the lows (Mann-Whitney-Wilcoxon test, two-sided $p_{high}<0.005$ and $p_{low}=0.9$, FDR-corrected; for the lows, $BF_{01}=4$).

Likewise, during AS, only high synchronizers significantly engaged the fronto-parietal network (Fig. 3A; Mann-Whitney-Wilcoxon test, two-sided $p_{high}<0.005$ and $p_{low}=0.42$, FDR-corrected; for the lows, $BF_{01}=4.1$), though its activity did not differentiate between the groups. Given that groups were defined by their speech auditory-motor synchrony, we then correlated the engagement of the fronto-parietal network with the synchronization (PLV) between the perceived and produced syllables during AS. Indeed, these measures were positively correlated across the whole sample as well as for the high synchronizers only (Fig. 3B; Spearman correlation coefficient $r_{all}=0.41$ and $r_{high}=0.56$, $p_{all}=0.009$ and $p_{high}=0.012$).

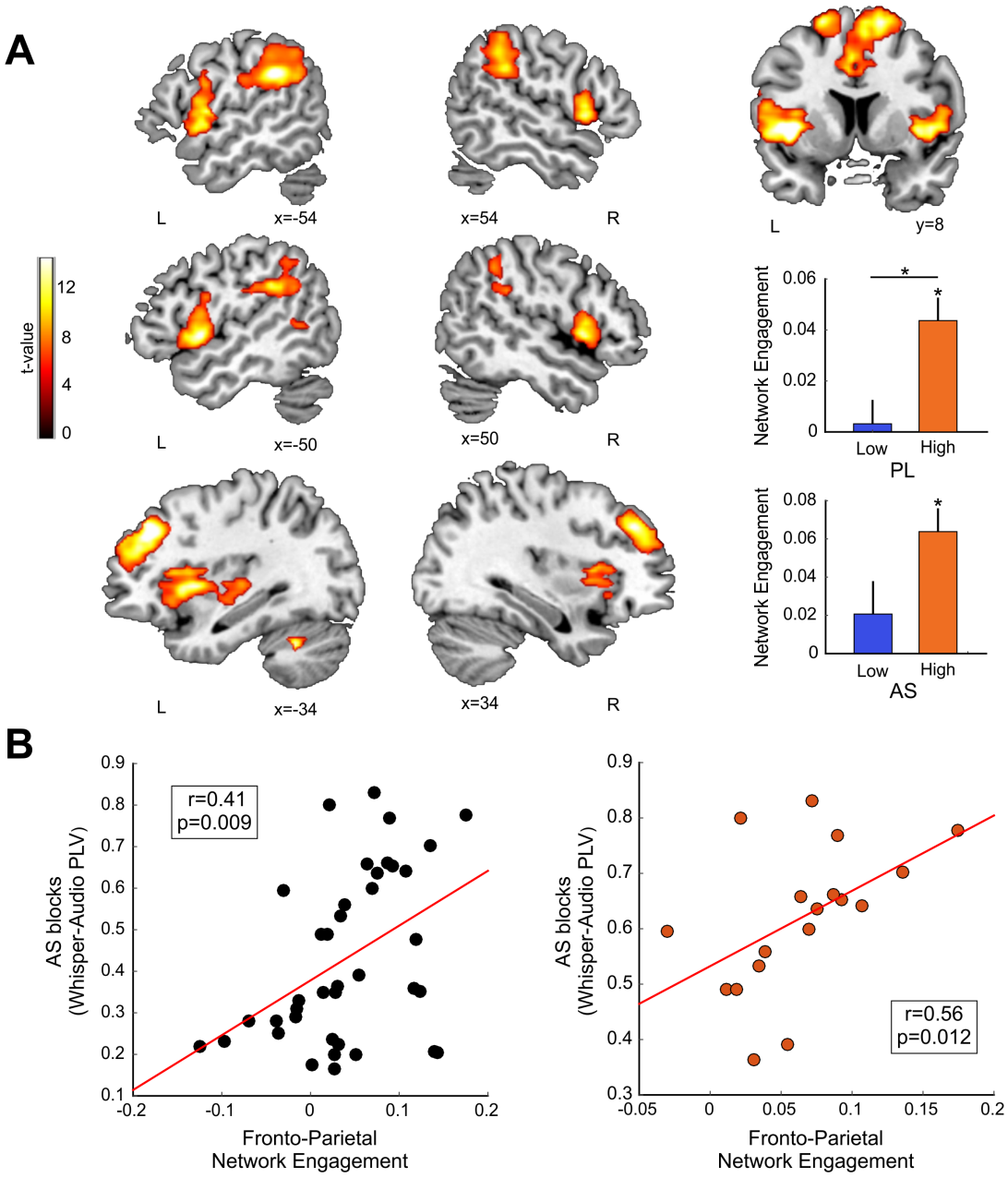


**Figure 3. High synchronizers activate an additional brain network during statistical word learning.** **(A)** In red/yellow, the fronto-parietal network is shown over a canonical template, with MNI coordinates on the lower portion of each slice. Neurological convention is used. A $p<0.05$ FWE-corrected threshold at the cluster level with an auxiliary $p<0.001$ threshold at the voxel are used. This is the only network showing significant differences in activity between high and low synchronizers during PL (see bar plots on the lower right; * $p<0.05$, FDR-corrected). (**B)** Scatterplot displaying participants' PLV during AS as a function of the fronto-parietal network's engagement. Red line represents the correlation of the data. *Left inset:* all subjects. *Right inset:* high synchronizers.

*Neuroimaging results (II): The interplay between networks boosts learning*

Next we assessed whether the activity of any of the networks significantly engaged during the PL condition was predictive of word learning. Replicating previous results[12,24], we found a network comprising mainly bilateral auditory regions and a small superior pre/motor cluster (henceforth, auditory network); whose activity positively correlated with learning performance in the whole sample (Fig. 4A; Spearman correlation coefficient r=0.42 and p=0.032, FDR-corrected). We found no significant correlations between learning and network activity in the AS condition. Since during PL: *i)* highs behaviorally outperformed lows; *ii)* the fronto-parietal network was only activated by high synchronizers; and *iii)* the auditory network was related to learning performance, we explored the relationship between the time courses of these two networks at the individual level. As illustrated in Figure 4B, high synchronizers with a greater learning benefit (defined as PL-AS) appeared to show a distinct pattern with interweaving time courses between auditory and fronto-parietal networks. To quantify this observation, we employed an analysis typically used in electronics: XOR. Applied to our signals, this logical operation assigns a single value per time point: *one* (1) when a single network is above baseline activity, or *zero* (0) otherwise; that is, a *one* is assigned when one or the other network (but not both) is active (Fig. 4B, lower insets). For each high synchronizer, we averaged the XOR over time, and correlated this value with their learning benefit (PL-AS) (note that this analysis would be meaningless for low synchronizers, given the non-significant activation of their fronto-parietal network). In line with the observed pattern, a positive correlation was found (Fig. 4C; Spearman correlation coefficient r=0.65, p<0.005). This suggests that the learning benefit shown by high synchronizers over lows was driven by a complex interplay between fronto-parietal and auditory networks.

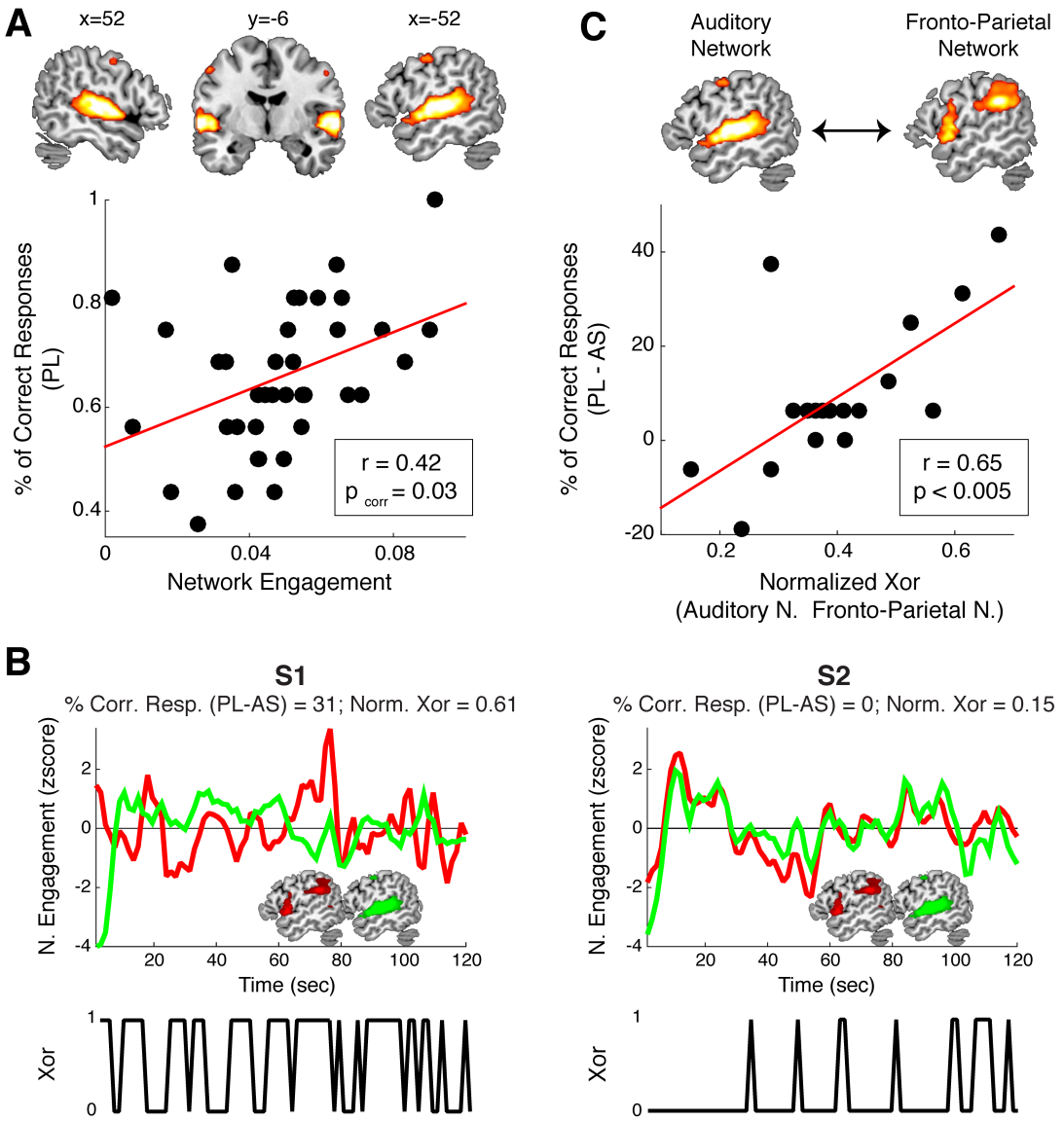


**Figure 4. An interplay between networks boosts learning. (A)** The auditory network supports learning during PL. *Upper inset:* In red yellow, the auditory network is shown over a canonical template, with MNI coordinates on the upper portion of each slice. Neurological convention is used with a $p<0.05$ FWE-corrected threshold at the cluster level, with an auxiliary $p<0.001$ threshold at the voxel level. *Lower inset:* Scatterplot displaying participants' percentage of correct responses during PL as a function of the auditory network's engagement. **(B)** The learning benefit during PL (correct answers in PL − correct answers in AS) is related to the interplay between the time courses of the fronto-parietal (red) and the auditory (green) networks. *Left/right inset:* a representative high synchronizer with a greater/smaller learning benefit. *Lower insets:* Time evolution of the XOR analysis. **(C)** Scatterplot displaying high synchronizers' learning benefit as a function of the normalized XOR between the fronto-parietal (red) and the auditory (green) networks. Red line: correlation of the data.

## DISCUSSION

Here we show that the neural substrates supporting SWFL vary across individuals in a systematic way. We arrive at this observation by splitting the population into two groups according to their spontaneous speech auditory-motor synchronization abilities. Specifically, we found two distinct networks related to SWFL. One, involving mainly auditory regions and a small superior pre/motor cluster (auditory network), is universally recruited and directly correlates with learning. Another, which includes inferior frontal, inferior parietal and supplementary motor areas (fronto-parietal network) is neither necessary nor sufficient for learning yet boosts performance. This latter network is exclusively recruited by high auditory-motor synchronizers. These observations thus parsimoniously account for the apparently disparate results in previous SWFL literature.

In terms of behavior, we find that the detrimental effects of AS are not universal. We demonstrate –in two independent samples– that only the performance of participants with a high degree of auditory-motor synchronization -i.e. high synchronizers- is affected by AS. Low synchronizers, in contrast, remain unaltered. These results indicate that, while speech production regions are not be necessary for SWFL, their additional recruitment confers a benefit in learning: high synchronizers, who do show AS effects, performed better than lows during PL. Note that these results are not discordant with the previous literature[9] since averaging high and low synchronizers yields the expected AS effects.

At the neural level, we found an important distinction between high and low synchronizers related to the engagement of a fronto-parietal network: only high synchronizers engage this network during PL. While SWFL performance correlates with the activity of the auditory network across the entire sample -in line with previous literature[12]- a synergistic relationship between both networks boosts learning in the high synchrony group. Additionally, the engagement of the fronto-parietal network predicted the degree of spontaneous synchronization of produced speech during the AS condition.

A relationship between auditory-motor synchronization and language skills has been previously reported in the literature[16,25,26]. For example, precision in tapping to a metronome correlates with reading and spelling abilities in children with developmental dyslexia[27]. Similarly, synchronization to a beat correlates with phonological awareness and rapid naming in typically-developing preschoolers[28]. Despite the cumulative evidence for a link between

auditory-motor synchronization and this various language skills, the existence of a neural substrate shared among these apparently unrelated cognitive abilities remains an empirical question. In this direction, our results suggest that the reported fronto-parietal cortical network subserves this shared role: On one hand, the engagement of this network during PL confers a benefit in learning; on the other, during AS, the activity of this network predicts the degree of speech auditory-motor synchronization.

From a mechanistic perspective, we propose that enhanced syllabic parsing -a prerequisite for SWFL- results from the coordinated activity between auditory and fronto-parietal networks, leading to a better statistical learning performance. In line with this, we previously showed that frontal-most component of the fronto-parietal network, the IFG, aligns with the onset of perceived syllables in high synchronizers. This frontal region has also been shown to send top-down signals to the auditory cortex (the main component of the auditory network) to better align its activity to the speech input[29,30]. A similar proposal has been advanced in the literature to account for the enhanced processing of phonology that results from improved auditory timing perception through auditory-motor training[31].

In sum, our work sheds light onto the neural substrates of SWFL by taking into account individual differences. We show that what appeared to be disparate results in the existing literature stems from pooling together fundamentally distinct populations. Specifically, we reveal that, beyond a universally recruited network for SWFL, a fronto-parietal network enabling auditory-motor synchronization is selectively engaged by some individuals to produce a benefit in learning. The SSS-test thus emerges as a useful tool to achieve a more nuanced characterization of speech related phenomena. In all, this work sounds a note of caution about assuming the existence of monolithic mechanisms underlying cognitive tasks.

# METHODS

## Participants

A first cohort of 65 participants completed the behavioral protocol. Nine participants were removed because they spoke loudly instead of whispering or because they stopped whispering for longer than 4 sec (during the SSS-test and/or the AS blocks). The data from 56 participants (34 females; mean age, 22; age range, 18 to 37) were analyzed.

A second cohort comprising 22 low and 22 high synchronizers participated in the fMRI protocol. Participants were selected from a larger group of 388 individuals, which completed the SSS-test in the context of a previous study[16] (Fig. S1A). Three participants were removed due to artifactual fMRI data. The final dataset comprised 41 individuals (23 females; mean age, 28; age range, 20 to 54)

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

## Overall experimental design

The behavioral protocol consisted of four blocks of statistical word-learning performed under two different conditions (PL and AS), followed by the SSS-test (see Fig. 5A). Four *pseudo-languages* were generated and presentation order was randomized across participants. During the PL condition participants passive listened to the *pseudo-languages*, while during AS they concurrently, and repeatedly, whispered the syllable *"tah"*. Conditions were interleaved and the two possible orders (PL-AS-PL-AS or AS-PL-AS-PL) were randomized across participants.

The behavioral protocol was modified for fMRI acquisition. First, we divided the protocol into two experimental runs with one AS and one PL block each. In addition, one minute of Resting state was introduced before each statistical word-learning block (PL or AS) and two minutes of speech production without auditory input (Speech motor condition) were introduced at the end of each run (see Fig. 5A). Specifically, the Speech motor condition consisted in repeatedly whispering the syllable *"tah"* with no auditory input. The SSS-test was not included in

the fMRI session. Participants' speech synchronization abilities were assessed in a previous study[16].

Importantly, participants' whisper was recorded during every AS block for both the behavioral and fMRI versions of the experiment (for the latter we used an MRI compatible, noise cancelling microphone, OptoAcoustics FOMRI).

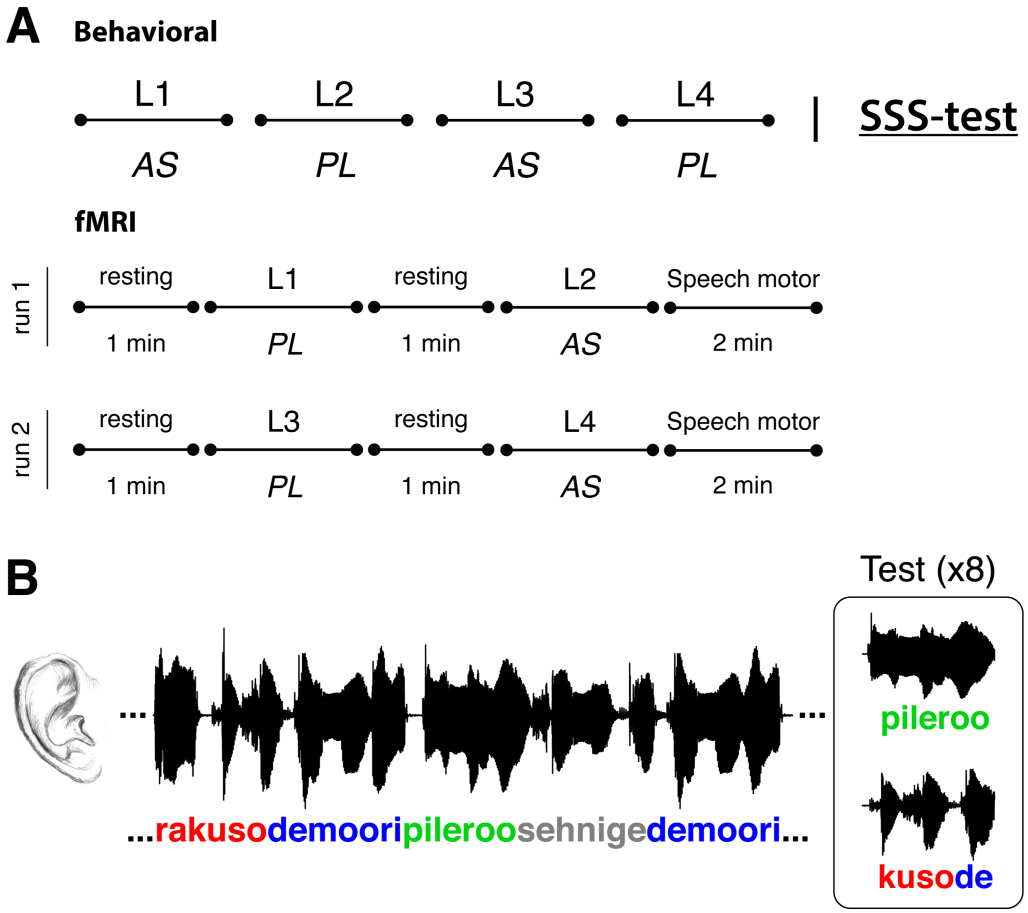


**Figure 5. Overall experimental design. (A)** The behavioral protocol consisted in 4 blocks of statistical word-learning (each comprised a different *pseudo-language*) followed by the SSS-test. The statistical word-learning blocks were completed under 2 different conditions: PL, wherein participants passively listened to the *pseudo-languages* and AS, where participants concurrently, and repeatedly, whispered the syllable *"tah*". Conditions were interleaved and the order (AS-PL-AS-PL or PL-AS-PL-AS) was counterbalanced across participants. *Lower inset:* for the fMRI session, 2 speech-motor and 4 resting state blocks were added to the behavioral protocol. **(B)** Schematic representation of a statistical word-learning block. *Left inset:* learning phase. Participants listened to the two-minute-long auditory stream containing the four words of the *pseudo-language*. *Right inset:* testing phase. Learning was assessed by 8 two-alternative forced choices between a *word,* upper line, and a *part-word,* lower line.

**Stimuli**

For the SSS-test and the word-learning task, we created five sets of stimuli (G1 to G5), each containing 12 distinct syllables (unique consonant-vowel combinations). Syllables for each set were handpicked to maximize both between and within set variability. G1 to G4 were used to make the stimuli for the word-learning task (i.e., word streams, words, and part-words), while G5 was exclusively used for the SSS-test. Four pseudo-languages (L1 to L4) were created by the combination of all the syllables in each set (G1 to G4, respectively) to form tri-syllabic pseudo-words (henceforth, words). Each pseudo-language thus contained four distinct words, relatively balanced on word-average bi-phoneme and positional probabilities according to The Irvine Phonotactic Online Dictionary (IPhOD version 2.0; http://www.IPhOD.com/), so as to maximize their learnability. Words were combined to form streams, with no gaps between words, lasting 2 minutes each. An equal number of non-consecutive repetitions per word was ensured. Part-words were created by the concatenation of a word's final syllable and the first two syllables of all the remaining words of the same pseudo-language (i.e., 12 part-words for each pseudo-language). Five independent native speakers produced written renderings of all words and part-words. Renderings with the highest convergence were selected for the visual presentation during the test phase of the word-learning task. A minute-long random syllable stream for the SSS-test was created by the random combination of the syllables in G5, with no pauses between them and ensuring no consecutive repetitions. Words, part-words and streams were converted to .wav files using the American Male Voice diphone database (US2) of the MBROLA text-to-speech synthesizer[32] at 16 kHz. All phonemes were equal in duration (111ms) - satisfying a constant syllable presentation rate of 4.5Hz, pitch (200Hz), and pitch rise and fall (with the maximum in the middle of the phoneme).

**Statistical Word-Learning Task**

The statistical word-learning task consisted of the following phases: a learning phase, during which participants listened to the 2-minute-long streams containing the four words of the *pseudo-language* (L1 to L4); and a test phase, where each word of the *pseudo-language* was presented against a part-word (randomly selected from the pool of 12) in a two-alternative forced choice (see Fig. 5B). During each test, words and selected part-words were presented twice each, within non-repeating pairs, making this a total of eight test trials. Test items were visually (on-screen) and auditory presented and were assigned a '1' or a '2' according to the order of auditory

presentation and left-to-right on-screen presentation. Participants were required, for each test pair, to indicate their choice by pressing the corresponding number. The presentation of the *pseudo-languages* was counterbalanced between participants. Group analyses were performed on the proportion of correct responses averaged across same-type conditions (PL and AS).

**SSS-test**

Participants in a sound isolated booth, seated in front of a PC with a microphone placed close to their mouth, listened to a rhythmic stream of syllables while concurrently and repeatedly whispering the syllable "*tah"* during one minute. Next, they indicated whether a given set of target syllables was presented. Participants were not explicitly instructed to synchronize their speech to the external audio; instead the instruction was to correctly recall the syllables. Before the test, participants were primed to repeatedly whisper *"tah"* at a rate of 4.5 syll/sec; the same rate at which the external syllables were presented. For further details about the test, see *Assaneo et al. 2019*[16].

**Speech synchronization measurement**

The degree of synchronization was measured by the phase locking value (PLV) between the envelope of the produced speech and the cochlear envelope[21,33] of the rhythmic syllable stream or the *pseudo language*. 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 $\theta_1$ and $\theta_2$ the phase of the first and the second signals, respectively. 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.

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.

**Definition of high and low synchronizers**

Each participant was classified as a low or a high synchronizer according to their corresponding speech synchronization value (Whisper to Audio PLV) obtained during the SSS-

test. A threshold value was estimated from a previous dataset[16] comprising 388 PLVs obtained with different versions of the SSS-test (Fig. S1A). We applied a *k-means* clustering algorithm[34], using a squared Euclidean distance metric with 2 clusters, and computed the midpoint between the clusters' centers ($PLV_{threshold}$=0.49). Participants with a PLV below/above this value were classified as low/high synchronizers.

**Scanning parameters**

All fMRI scans took place at the NYU Center for Brain Imaging using a 3T scanner (Siemens Prisma 3T MRI scanner) and a 64-channel phased-array head coil. For the statistical word-learning fMRI task, two runs of at least 400 (note that the test is self-paced and total timings vary between participants) sequential whole-brain multi-echo echo-planar imaging (EPI) volumes were acquired (TR = 1500 ms, TE=45 ms, flip angle = 77º, voxel size = 2.0 × 2.0 × 2.0 $mm^3$, 64 axial slices, acquisition size = 104×104). A high resolution T1 MPRAGE image was also acquired (TR = 2400 ms, TE = 2.24 ms, flip angle = 8º, voxel size = 0.80 × 0.80 × 0.80 $mm^3$, 256 sagittal slices, acquisition matrix = 320 × 300).

**fMRI and ICA preprocessing**

Data were preprocessed using MATLAB R2018a and the Statistical Parameter Mapping software (SPM12, Wellcome Trust Centre for Neuroimaging, University College, London, UK, www.fil.ion.ucl.ac.uk/spm/). For each subject, we first realigned the two word-learning runs to the mean image of all EPIs. The T1 was then co-registered to this mean functional image and segmented using Unified Segmentation[35]. The deformation fields obtained during the segmentation step were used to spatially normalize all functional images from each run to the MNI template included in SPM12 (we maintained the original acquisition voxel size of 2.0 × 2.0 × 2.0 $mm^3$). Images were finally spatially smoothed with a 6 mm FWHM kernel.

We used the Group ICA of fMRI Toolbox[19] (GIFT v4.0b; http://mialab.mrn.org/software/gift/) to apply group spatial ICA to the previously preprocessed fMRI data. Based on previous research in clinical and healthy populations[12,36,37], the number of independent components to be extracted was set to 20. Data were intensity normalized, concatenated and, using principal component analysis, reduced to 20 temporal dimensions. Then, this preprocessed data were fed to the *infomax* algorithm[38]. The intensities of the spatial maps

were in percentage of signal change after the intensity normalization, and thus no scaling was used.

In order to assess which of the 20 ICA networks retrieved were related to the different conditions of interest (PL and AS), both spatial and temporal classification methods were employed. First, for all subjects, the spatial map of each individual ICA network was submitted to a second-level analysis using a one sample *t*-test under SPM12[39]. We then obtained, for each network, a group map of activity that was thresholded using a $p < 0.05$ FWE-corrected threshold at the cluster level, with an auxiliary $p < 0.001$ threshold at the voxel level. Clusters with fewer than 50 voxels were not included in the analyses. We visually inspected these thresholded networks and 8 were discarded after being considered artefactual (see Results).

Using GIFT, for the remaining 12 networks, we calculated a multiple regression that fitted each participant's network time-course to a model. The model was created using SPM12, by convolving the timing of both the main (AS and PL) and control (Rest and Speech motor) conditions with a canonical hemodynamic response. In order to control for motion artifacts, the model also included six movement regressors obtained from the realignment step. By fitting a multiple regression between this model and each network's time-course, we obtained, for each condition, beta values that represented network engagement. For PL beta values, we used the Rest condition as a baseline. For AS we used both the Rest and the Speech motor control conditions as baseline, in order to capture the activity related to the learning process during AS itself and not to the motor-activity related to the whispering. For any comparison using beta values, participants exceeding 2 SD were excluded from the analysis.

For the group spatial maps of each network, maxima and all coordinates are reported in MNI space. Anatomical and cytoarchitectonical areas were identified using the Automated Anatomical Labeling[40] and the Talairach Daemon[41] database atlases included in the xjView toolbox (http://www.alivelearn.net/xjview/).

**Statistical Analyses**

Throughout this manuscript, we used non-parametric Mann-Whitney-Wilcoxon and Wilcoxon signed-rank tests for between and within subject comparisons, respectively. Multiple comparisons were controlled using a False-Discovery Rate correction. Non-parametric Spearman's rank correlations were used to assess the relationship between variables. Bayes

factors ($BF_{01}$), which reflect how likely data are to arise from the null model (i.e., the probability of the data given H0 relative to H1), were also computed with the software JASP using default priors[42–44].

**Acknowledgements**

This work was supported by NIH grant 2R01DC05660 (D.P.) and FP7 Ideas: European Research Council grant ERC-StG-313841 (R.d.D.B.).

**Author contributions**

M.F.A., P.R., J.O., D.L.B, R.d.D.B. and D.P. designed the research and wrote the manuscript; M.F.A., P.R., J.O. and L.N. acquired and analyzed the data.

**Competing interests**

The authors declare no competing interests.

**Materials & Correspondence**

Correspondence and material requests should be addressed to M. Florencia Assaneo.

**References**


1. Gervain *et al.* The neonate brain detects speech structure. *Proc. Natl. Acad. Sci.* (2008). doi:10.1073/pnas.0806530105

2. Bulf *et al.* Visual statistical learning in the newborn infant. *Cognition* (2011). doi:10.1016/j.cognition.2011.06.010

3. Saffran *et al.* Statistical learning by 8-month-old infants. *Science (80-. ).* **274,** 1926–1928 (1996).

4. Saffran *et al.* Statistical learning of tone sequences by human infants and adults. *Cognition* (1999). doi:10.1016/S0010-0277(98)00075-4

5. Fiser & Aslin. Statistical Learning of Higher-Order Temporal Structure from Visual Shape Sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* (2002). doi:10.1037/0278-7393.28.3.458

6. Conway & Christiansen. Modality-constrained statistical learning of tactile, visual, and auditory sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* (2005). doi:10.1037/0278-7393.31.1.24

7. Toro & Trobalón. Statistical computations over a speech stream in a rodent. *Perception and Psychophysics* (2005). doi:10.3758/BF03193539

8. Meyer & Olson. Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc. Natl. Acad. Sci.* (2011). doi:10.1073/pnas.1112895108

9. Lopez-Barroso *et al.* Language learning under working memory constraints correlates with microstructural differences in the ventral language pathway. *Cereb. Cortex* **21,** 2742–2750 (2011).

10. McNealy *et al.* Cracking the Language Code: Neural Mechanisms Underlying Speech Parsing. *J. Neurosci.* (2006). doi:10.1523/jneurosci.5501-05.2006

11. Cunillera *et al.* Time course and functional neuroanatomy of speech segmentation in adults. *Neuroimage* (2009). doi:10.1016/j.neuroimage.2009.06.069

12. López-Barroso *et al.* Multiple brain networks underpinning word learning from fluent speech revealed by independent component analysis. *Neuroimage* (2015). doi:10.1016/j.neuroimage.2014.12.085

13. Karuza *et al.* The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain Lang.* (2013). doi:10.1016/j.bandl.2012.11.007

14. Henin *et al.* Statistical learning shapes neural sequence representations. *bioRxiv* 583856 (2019). doi:10.1101/583856

15. Lopez-Barroso *et al.* Word learning is mediated by the left arcuate fasciculus. *Proc. Natl. Acad. Sci.* **110,** 13168–13173 (2013).

16. Assaneo *et al.* Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. *Nat. Neurosci.* (2019). doi:10.1038/s41593-019-0353-z

17. Hickok & Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience* **8,** 393–402 (2007).

18. Baddeley *et al.* Exploring the Articulatory Loop. *Q. J. Exp. Psychol. Sect. A* (1984). doi:10.1080/14640748408402157

19. Calhoun *et al.* A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* **14,** 140–151 (2001).

20. Pellegrino *et al.* Across-Language Perspective on Speech Information Rate. *Language (Baltim).* **87,** 539–558 (2011).

21. Ding *et al.* Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81,** 181–187 (2017).

22. Varnet *et al.* A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* **142,** 1976–1989 (2017).

23. Assaneo & Poeppel. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Sci. Adv.* **4,** eaao3842 (2018).

24. Plante *et al.* The nature of the language input affects brain activation during learning from a natural language. *J. Neurolinguistics* (2015). doi:10.1016/j.jneuroling.2015.04.005

25. Thomson *et al.* Auditory and motor rhythm awareness in adults with dyslexia. *J. Res. Read.* (2006). doi:10.1111/j.1467-9817.2006.00312.x

26. Tierney & Kraus. The ability to tap to a beat relates to cognitive, linguistic, and perceptual

skills. *Brain Lang.* (2013). doi:10.1016/j.bandl.2012.12.014

27. Thomson & Goswami. Rhythmic processing in children with developmental dyslexia: Auditory and motor rhythms link to reading and spelling. *J. Physiol. Paris* (2008). doi:10.1016/j.jphysparis.2008.03.007

28. Woodruff Carr *et al.* Beat synchronization predicts neural speech encoding and reading readiness in preschoolers. *Proc. Natl. Acad. Sci.* **111,** 14559–14564 (2014).

29. Rimmele *et al.* Proactive Sensing of Periodic and Aperiodic Auditory Patterns. *Trends in Cognitive Sciences* (2018). doi:10.1016/j.tics.2018.08.003

30. Park *et al.* Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners. *Curr. Biol.* **25,** 1649–1653 (2015).

31. Tierney & Kraus. Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (PATH). *Front. Hum. Neurosci.* (2014). doi:10.3389/fnhum.2014.00949

32. Dutoit & Pagel. 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).

33. Chi T. NSL Matlab Toolbox. *Maryl. Neural Syst. Lab., Univ. Maryland.*

34. Arthur & Vassilvitskii. k-means++: The Advantages of Careful Seeding. *Proc. eighteenth Annu. ACM-SIAM Symp. Discret. algorithms* 1–11 (2007). doi:10.1145/1283383.1283494

35. Ashburner & Friston. Unified segmentation. *Neuroimage* (2005). doi:10.1016/j.neuroimage.2005.02.018

36. Sihvonen *et al.* Functional neural changes associated with acquired amusia across different stages of recovery after stroke. *Sci. Rep.* (2017). doi:10.1038/s41598-017-11841-6

37. Smith *et al.* Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci.* (2009). doi:10.1073/pnas.0905267106

38. Bell & Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Comput.* **7,** 1129–1159 (1995).

39. Calhoun *et al.* fMRI activation in a visual-perception task: Network of areas detected using the general linear model and independent components analysis. *Neuroimage* (2001). doi:10.1006/nimg.2001.0921

40. Tzourio-Mazoyer *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* (2002). doi:10.1006/nimg.2001.0978

41. Lancaster *et al.* Automated Talairach Atlas labels for functional brain mapping. *Hum. Brain Mapp.* (2000). doi:10.1002/1097-0193(200007)10:3<120::AID-HBM30>3.0.CO;2-8

42. JASP Team. JASP (Version 0.9)[Computer software]. *[Computer software].* (2018).

43. Morey *et al.* Package 'BayesFactor'. *Compr. R Arch. Netw.* (2018).

44. Rouder & Morey. Default Bayes Factors for Model Selection in Regression. *Multivariate Behav. Res.* (2012). doi:10.1080/00273171.2012.734737