Speech imagery decoding as a window to speech planning and production

Joan Orpella[1, *], Francesco Mantegna[1], M. Florencia Assaneo[2], David Poeppel[1,3,4]

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

[2]Institute of Neurobiology, National Autonomous University of Mexico, Juriquilla, Querétaro, Mexico.

[3]Center for Language, Music and Emotion (CLaME), New York University, New York, NY, USA and Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany.

[4]ESI, Frankfurt, Germany.

* Corresponding author (jo1358@nyu.edu)

Abstract

Speech imagery (the ability to generate internally quasi-perceptual experiences of speech events) is a fundamental ability tightly linked to important cognitive functions such as inner speech, phonological working memory, and predictive processing. Speech imagery is also considered an ideal tool to test theories of overt speech. Despite its pervasive nature, the study and use of speech imagery for clinical or basic research has been tremendously challenging, primarily because of the lack of behavioral outputs and the difficulty in temporally aligning imagery events across trials and individuals. Here we used magnetoencephalography (MEG) paired with time-resolved decoding and a novel behavioral protocol to map out the processing stages underlying speech imagery. We monitored participants' upper lip and jaw micromovements during imagery using electromyography. Decoding of participants' imagined syllables revealed a rapid sequence of representations from visual encoding to the imagined speech event. Importantly, participants' micromovements did not discriminate between the syllables. The neural correlates of the decoded sequence maps neatly onto the predictions of current computational models of speech motor control and provide some evidence for hypothesized internal and external feedback loops for speech planning and production, respectively. Additionally, a windowed multinomial classification (WMC) analysis revealed the presence of two nested and concurrent levels of representation (syllable and consonant-vowel transition) and the compressed nature of representations during planning. It is assumed that the same sequence underlies the motor-based generation of sensory predictions that modulate speech perception and the articulatory loop of phonological working memory. The results highlight the potential of speech imagery for different research domains, based on these new experimental approaches and analytical methods, and further pave the way for successful non-invasive brain-computer interfaces.

Introduction

Mental imagery of speech refers to the internally generated quasi-perceptual experience of our own or others' speech. Research on speech imagery has a long and well-established history in the sciences and in philosophy before that[1–5], and there are many reasons for this longstanding interest. For one, speech imagery lies at the core of critical cognitive functions, such as inner speech and phonological working memory, which have important implications for learning, problem-solving, and, more generally, development[6,7]. Speech imagery is also considered an adequate model for overt speech[7–10]. As such, it has been employed in research to gain insights into aspects of speech planning, production, and motor control otherwise difficult to test with overt speech[8,11–16]. Imagery (speech or other) is moreover a paradigmatic example of the generation of sensory predictions[8,17,18]. In speech, the hypothesis is that 'that little voice in our head' results from an internal prediction of the sensory consequences of planned motor commands[19], that is, from some sort of internal emulation[20]. Critically, these predictions can be used to anticipate sensory inputs, such as other's speech, which facilitates comprehension[8]. This makes speech imagery a potentially invaluable tool to test predictive processing theories of the mind[21], such as predictive coding[22,23], Bayesian inference[24], and associative learning[25–29]. Finally, speech imagery is also clinically relevant. Imbalances between sensory predictions and feedback are thought to underly disorders such as schizophrenia, autism, and stuttering[6,7,30]. Moreover, advances in the decoding of speech imagery are potentially life-changing for individuals that have lost the ability to speak due to stroke or illness[31–35], besides their many other conceivable applications in industry.

Despite its potential as a research and clinical tool and the success of imagery research in other domains (e.g., limb motor control[36]), speech imagery remains poorly characterized. This is mostly due to methodological challenges[6,37] (e.g., the lack of comparative research, the lack of behavioral outputs, the potential misalignments across experimental trials and participants) and the lack of appropriate paradigms and analytical approaches to overcome them. Notwithstanding, researchers have used speech imagery in ingenious ways, for example, to test or localize specific aspects of speech motor control (e.g., feedback prediction errors[13]), but evidence remains indirect and fractional. For instance, the experimental modulation of speaking-induced suppression and its effects on perception can only *imply* the existence of so-called forward models and of precise sensory predictions emanating from planned speech[13,38,39]. The same methodological challenges permeate the many attempts to decode speech imagery with time-resolved methods[37], which have only recently began to produce some hopeful results albeit restricted to state-of-the-art invasive (intracranial) recordings and highly sophisticated analysis pipelines (e.g., [40]).

Here we capitalized on the excellent balance between temporal and spatial resolution of a non-invasive method, magnetoencephalography (MEG), paired with a deceptively simple speech imagery task (Fig 1) and a powerful decoding approach[41,42] to map out the sequence of neural processes underlying speech imagery. In short, we decoded participants´ imagined speech as it unfolds.

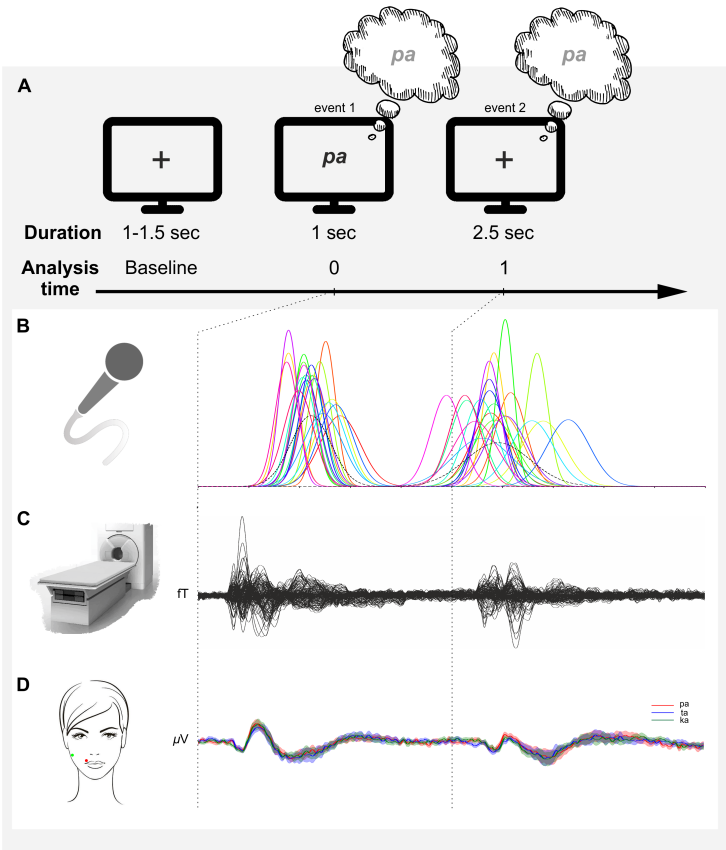

Fig 1. **Experimental protocol**. **A. Task**. Each trial began with a fixation cross of variable duration (1-1.5 sec) in the center of the screen. One of three syllables (*pa*, *ta*, or *ka*) was then presented and remained on the screen for exactly 1 second. Syllable presentation was followed by another fixation cross lasting 2.5 seconds, after which the next trial began. The experiment comprised two conditions, *Imagery* and *Reading*, each with 4 blocks of 120 trials (40 presentations of each syllable per block), counterbalanced between participants. The total number of trials over the course of the experiment was therefore 960 (320 per syllable). Syllable presentation was fully randomized within each block. In the *Imagery* blocks, participants were instructed to imagine producing the given syllable as soon as possible after its appearance on the screen (*event 1*) and a second time on the fixation cross (*event 2*). In the *Reading* blocks, participants were instructed to passively look at the center of the screen. Prior to the MEG session (1-7 days), participants received a training session in which the experimenter explained the task and the desired type of imagery (i.e., imagining producing vs simply hearing). They were also asked to complete a full block using *overt* productions of the syllables (aloud rather than imagined). At this time only participants were given feedback regarding the timings of each production. This was critical in ensuring a minimum temporal alignment within participant as well as consensus across the cohort. Additionally, participants were also given a link to an online version of the task (available here) to practice with in their own time. **B. Expected time of imagery.** On the day of the MEG acquisition, each participant completed a minimum of 1 practice block using overt productions, which were recorded for subsequent analysis. The figure shows all participants' syllable onset distributions for events 1 and 2. The medians of these distributions were important as for reference to the times when imagery was to be expected. **C. Average MEG data for a participant's *Imagery* trials**. Participants' neural activity was recorded during both *Imagery* and *Reading* trials using MEG. We used these data to decode participants' imagined syllables and map out the sequence of neural processes underlying imagery (see below). **D. Average electromyographic data for a participant's *Imagery* trials**. To monitor participants' movements during *Imagery*, we also recorded muscle activity from the upper lip (red dot) and jaw (green dot) using a MEG-compatible electromyography system. The figure shows expected micromovements during imagery which, critically, do not differ between the imagined syllables (see Fig S2 – S4 for the full analysis).

Specifically, we recorded MEG signals from 21 participants while imagining (internally producing) isolated syllables (*pa*, *ta*, and *ka*) prompted on the screen on each trial. We used syllables as the targets of speech imagery given recent evidence for syllable-like 'chunks' as fundamental units for speech perception and production ([43] for a review). First, we evaluated the extent to which these signals contained information over and above a *Reading* condition identical but for the instruction to internally produce the prompted syllable. Having established robust differences between the two conditions, we then asked whether *Imagery* trials contained decodable content regarding the actual syllables imagined, that is, whether the syllables that participants imagine can be decoded from their MEG data. Based on our decoding results, we were next able to map out the entire genesis and development of the imagined speech events, a sequence which had so far remained elusive to research. We then examined the dynamics of this sequence in order to further our understanding of how inner speech and sensory predictions are generated and to potentially adjudicate between current models of speech production, such as state feedback control[8,16,24,44,45] and DIVA[46]. Two main features distinguish SFC from other influential models, including DIVA. One is the existence of both an internal and an external feedback loop for speech planning and error monitoring, respectively (Fig 2). Another is the hierarchical organization of motor control units, with (roughly) syllabic and phonemic levels of control. Although

compelling evidence exists for some of these features in other domains of motor control[47], there is no direct evidence for them in speech. Interestingly, speech imagery was recently proposed as an ideal medium to investigate SFC[8,11]. To examine the question of internal and external feedback loops for internal speech planning and production, we assessed the time courses of auditory and motor areas during imagery. Finally, we enquired into the levels of motor control by probing the length of the representations involved in speech imagery via a novel decoding approach (windowed multinomial classification; WMC).

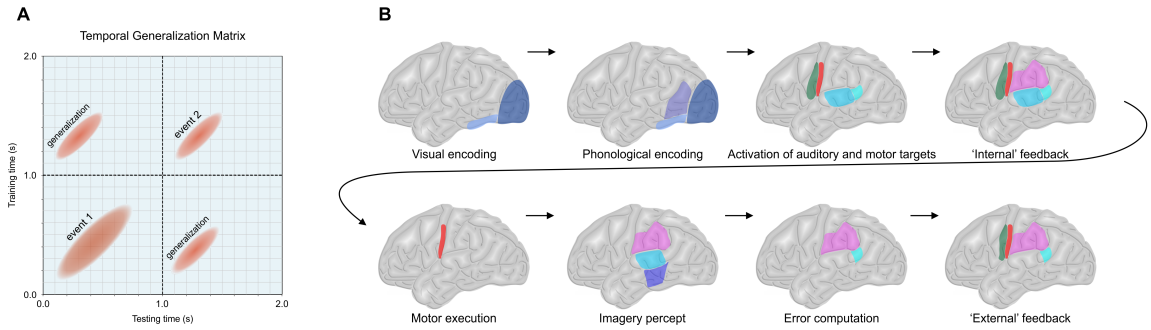

Fig 2. Main experimental hypotheses. A. Temporal Generalization method. To assess the presence of imagery and syllable decodeability, we used a multivariate pattern analysis in which classifiers are trained on a single time sample (rows; *y*-axis) but tested on all time samples of the trial (columns; *x*-axis). This results in a temporal generalization (TG) matrix that depicts the extent to which a given neural pattern is present across time. The TG method is a powerful approach to reveal not only the number and approximate times of neural processes but also the nature of the underlying representations (e.g., evolving, reactivated, ramping; see [41,42] for a full explanation of the method). We expected high decoding accuracy (red ellipses) in the extent to which neural processes underlying speech imagery juggle decodable representations. High accuracy was expected for both imagery events (event 1 and event 2). We also expected event 1 representations to generalize to event 2 and vice versa. **B. Expected sequence of neural processes for speech imagery.** Insofar as speech imagery mirrors overt speech, we hypothesized that the underlying sequence of neural processes to conform to current models of speech production. Inspired in a recent model[8,16,44] derived from other areas of motor control, we predicted 1. A speech planning stage encompassing visual encoding (visual and inferotemporal cortex), phonological encoding (left posterior middle and superior temporal cortex), the parallel activation of auditory and motor targets (auditory and motor cortex, respectively), and an 'internal' feedback loop for error correction characterized by sensory-motor activity; and 2. A production stage involving motor execution (motor cortex), activation of the imagery percept (auditory and somatosensory areas), prediction error elicited by the lack of overt auditory feedback (posterior auditory cortex), and a second feedback stage ('external' feedback) again with sensory-motor interactions. Note, that since we are assessing *imagery* rather than overt speech, the common nomenclature of *internal* and *external* loops, which refers to the nature of the feedback, is not strictly applicable. However, we decided to keep this nomenclature to link our predictions and potential results to theoretical formulations of the model.

Results

We recorded MEG signals from 21 participants (15 women; mean age = 28.19; *std* = 6.57) while they imagined producing one of three syllables (*pa*, *ta*, or *ka*). On every trial, participants were required to internally produce a given syllable as soon as it appeared on the screen (event 1) and a second time on a fixation cross appearing exactly 1 second later (event 2) (Fig 1). Participants electromyographic (EMG) data from the upper lip and jaw (Fig 1) was also acquired to measure any micromovements participants make during the MEG recordings. Although we expected micromovements during *Imagery*[48], in line with previous research[7], we did not expect these to discriminate between the different syllables. This is critical for the validity of our syllable decoding results from the MEG data.

Participants' performance on the overt version of the task is summarized in Table S1 (see also Fig S1). On average, the sound onset of syllable 1 (imagined event 1) occurred 439ms after the presentation of the syllable on the screen (*Methods*). The onset for syllable 2 occurred on average 175ms after the fixation cross. These times were taken to indicate the *expected time for imagery*, under the assumption of similar timings during the *Imagery* condition. Importantly, the interquartile range for the two events (syllable 1: 99ms; syllable 2: 146ms) indicated that participants were much more precise in time in the

production of the first imagined event than in the second. We expected these differences in variability to have an impact of decoding, with a greater alignment (event 1) translating into better decoding.

Since, by definition, there are no overt behavioral outputs of imagery, a first step was to measure and quantify the difference between the *Imagery* condition and a condition matched in all respects but for the instruction to imagine the syllables. In this *Reading* condition, participants simply looked at the syllables and fixation crosses appearing at the center of the screen. To evaluate the extent to which *Imagery* contained information over and above *Reading*, we used a decoding approach (temporal generalization; Fig 2). In addition to quantifying existing differences between two given conditions, this approach can potentially track the dynamics of neural processes underlying a particular experimental condition[41,42]. To track the development of neural processes common to speech imagery (i.e., processes shared by the three imagined syllables) as distinct from *Reading*, we trained a linear classifier on the *Imagery* vs *Reading* contrast at each time point and tested its performance across all timepoints within the trial. This analysis was performed for each subject separately using stratified 4-fold cross-validation with regularization and Receiver Operative Curve Area Under the Curve (ROC AUC) as a scoring metric (*Methods*). The analysis resulted in a temporal generalization (TG) matrix per subject, which we then averaged across subjects. We expected areas of higher decoding accuracy not only at the expected time of imagery (as determined from participants' overt productions during training; Fig S1; Fig 3A upper panel; Table S1) but also before imagined event 1, indicating motor planning for the syllable to be produced in that trial (Fig 2).

Fig 3A shows the average TG matrix across the entire sample (N = 21). Clusters of statistically significant decoding ($p < 0.05$; black contour lines) were determined at the second level of analysis via a cluster-based permutation test (1000 permutations; two-tailed) across subjects. Large areas of high decoding accuracy (with ROC AUC values up to 0.82) suggest robust and consistent differences over time between *Imagery* and *Reading* conditions. The extended nature of the clusters suggests that the differences are driven both by domain-general processes, such as attention, as well as processes more specific to speech imagery (i.e., bearing content). As expected, a significant degree of generalization can also be observed between the two imagery events within the trial, indicating similar neural underpinnings.

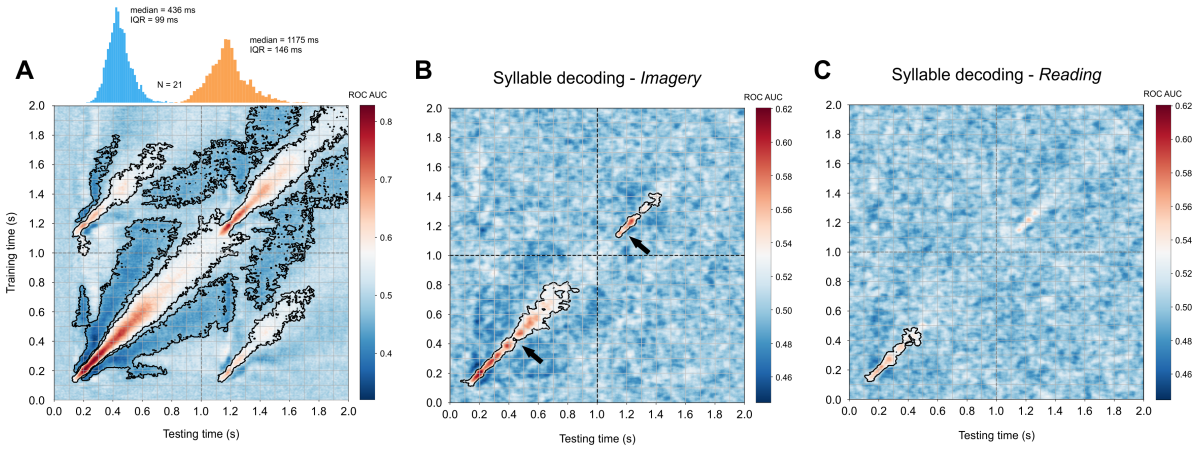

Fig 3. Temporal generalization matrices track the development of neural representations during speech *Imagery* and *Reading*. A. Average TG matrix (N = 21) for the contrast *Imagery* vs *Reading* and syllable onset time distributions (top inset) for all syllables overtly produced during training by all participants (blue = event 1; orange = event 2) plotted for reference as to the expected time for imagery. IQR = interquartile range. ROC AUC = Receiver Operative Curve Area Under the Curve (chance = 0.5). **B.** and **C.** TG matrices for the pairwise contrasts between syllables (*pa* vs. *ka, pa* vs. *ta,* and *ta* vs. *ka)* for each condition (*Imagery* and *Reading*, respectively*)* first averaged within subject and then across subjects (N = 21). Black arrows indicate the median syllable onset time of participants' overt productions during training (event 1 median: ~436ms; event 2 median: ~1175ms). Clusters of statistically significant decoding ($p < 0.05$; black contour lines) were in all cases determined at the second level of analysis via a cluster-based permutation test across subjects (1000 permutations; two-tailed). Statistical significance indicates consistence across subjects, while high ROC AUC values reflect robust classifier performance on discriminating the contrasts.

Decoding participants' imagined speech from the MEG data

Having established that the MEG signals for the *Imagery* condition contain information distinct from *Reading,* we next asked whether *Imagery* trials carried discriminable *content* regarding the three imagined syllables (*pa*, *ta*, and *ka*). Here, the TG approach can not only provide *direct* evidence for speech imagery if areas of significant decoding are found but also valuable insights into the nature of the neural processes involved (e.g., number of distinct processes, times of occurrence, generalizability). We first generated, for each participant and condition, a TG matrix for each of the pairwise syllable contrasts (*pa* vs. *ka, pa* vs. *ta,* and *ta* vs. *ka)* following the same decoding approach as before. These three matrices per participant and condition were then averaged within subject and finally entered into a cluster-based permutation test across subjects (N = 21; 1000 permutations; two-tailed) to determine clusters of significant syllable decodeability ($p < 0.05$) in each of the conditions (*Methods*). Given that the processing of visual information (i.e., reading the syllables) is shared between *Imagery and Reading*, we expected this analysis to yield some similarities in the early stages of each imagined event. However, direct evidence for speech imagery would also require for syllable decodeability to extend further in the *Imagery* condition only, reflecting the occurrence of the actual speech imagery event.

Islands of relatively high decoding accuracy (ROC AUC scores up to 0.62, significant at $p < 0.05$) during event 1 reveal a distinct cascade of neural processes during *Imagery* (Fig 3B). The limited span of these successive islands both on and off diagonal indicates that the representations involved were rapidly evolving (50ms-60ms) and highly specific (limited generalization). Significant decoding starts immediately after syllable presentation (~120ms) and extends well beyond the expected imagery time (Fig 3B black arrow at ~436ms). Syllable decodeability during event 2 was weaker albeit significant in clusters immediately before and after the expected imagery time (Fig 3B black arrow at ~1175ms). The weaker decoding for this event and the null generalization between events likely reflect the aforementioned misalignments within and between participants' inner productions as suggested by their overt productions (Fig 1; Fig 3A). As expected, syllable decodeability in the *Reading* condition was significant if weak in self-contained clusters between ~120ms and ~450ms (Fig 3C). This suggests a similar succession of neural processes to the *Imagery* condition up to the imagined event or extended visual processing during *Reading*. In favor of the latter interpretation, no clusters of significant decoding were found during event 2 in the *Reading* condition when no visual information of the syllable was present.

The analysis of the EMG data by participant indicated, as expected, the presence of micromovements (Fig S2 - Fig S4). Micromovements are a common phenomenon during imagery and inner speech and are commonly assumed to be a byproduct of motor signals that cannot be fully inhibited[7]. Interestingly, micromovements were present both in the *Imagery* and *Reading* conditions. We performed an in-depth analysis of the EMG data to ensure our decoding results could not be explained by participants' micromovements. Although small differences were found between *Imagery* and *Reading* conditions (Fig S2 and Fig S4), the micromovements did in no case discriminate between the imagined syllables (Fig S2 and Fig S4).

Neural dynamics underlying speech imagery

So far, we were able to decode participants' speech imagery and uncover a series of well-defined stages leading to the imagined event. We next sought to establish the neural correlates of these stages. On the one hand, this analysis can adjudicate between theories of speech imagery that posit a close parallel between imagery and overt production and theories that conceptualize imagery as a byproduct of motor planning (i.e., without primary motor involvement)[49]. On the other hand, if imagery mirrors overt speech, the sequence of neural events underlying imagery can adjudicate between current models of speech planning and production (Fig 2B).

To map out the neural dynamics underlying speech imagery, we first acquired structural MRI data from a random subsample of participants (*Methods*). Each participant's *Imagery* condition's average time

series was projected to their native source space and morphed to a common coordinate space (Montreal Neurological Institute) before averaging across participants (*Methods*). The goal of this group analysis was therefore to assess the sequence of neural activity that gives rise to speech imagery.

Fig 4 shows the progression of neural activity during *Imagery* between 120ms and 610ms after syllable presentation corresponding to the islands of significant syllable decoding previously identified. The analysis was thus guided by the decoding results (Fig 3B). Activity in source space is plotted alongside the corresponding sensor space topographies of the entire cohort. The close similarity between the larger cohort's topographies and both the sensor space topographies of the MRI sample and the classifier patterns (coefficients) for syllable decoding (cosine similarity tests Fig S5) indicates that the source reconstruction for the subsample is directly relevant to imagery.

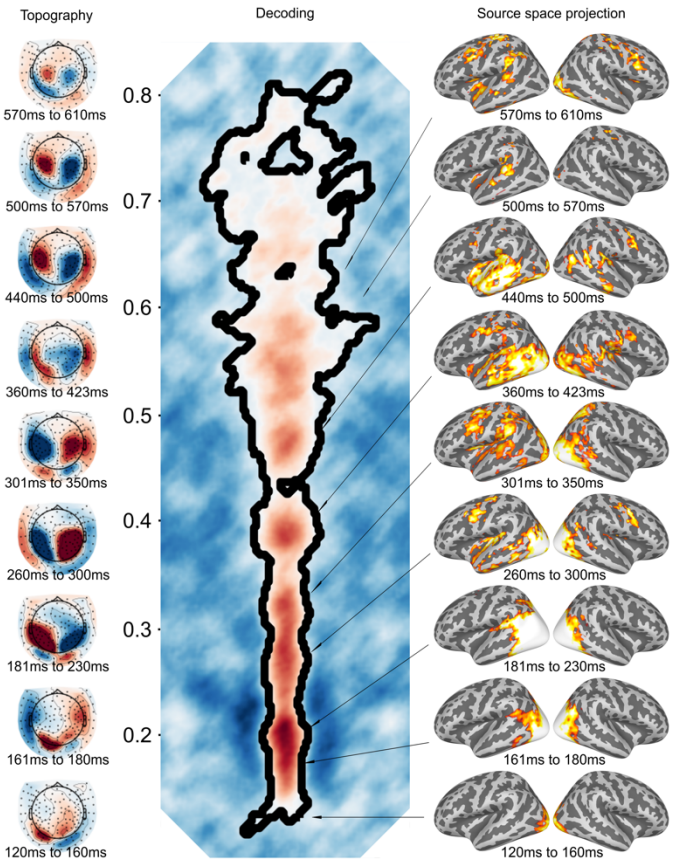

Fig 4. Clusters of syllable decodeability reveal a processing cascade during speech *Imagery*. Evoked response topographies (*Imagery*) averaged over participants (left panel) and evoked activity for a subgroup of participants estimated with sLORETA (right panel) corresponding to the clusters of significant syllable decoding in the *Imagery* condition, event 1 (center panel; from Fig 3). Evoked response topographies were thresholded at +-20fT. Source space activity was thresholded at minima ranging between 1.88 and 3.34 and maxima between 2.37 and 4.51 units for display purposes.

Both source and sensor renderings show a clear sequence of distinct neural events starting with visual areas (~120ms). This activity extends rapidly to well-established ventral and dorsal visual pathways and subsequently to the lateral temporal cortex (~180ms), particularly in the left hemisphere, coinciding both in location and time with a hypothesized phonological encoding stage[50,51]. This stage is followed by activity in the auditory cortex, area Sylvian parieto-temporal, frontal anterior insula, and bilateral pre-motor regions between 260ms and 300ms, potentially reflecting auditory-motor integration processes and speech motor planning[16,46,51,52]. Extensive activity over (predominantly left) auditory regions can be seen after 440ms, that is, at the expected time of imagery (estimated at 436ms for the entire sample and at 444ms for the MRI group; Table S1). We conjecture this activity to be the neural correlate of the quasi-perceptual experience that defines speech imagery, in line with previous research[11,15,19]. The decoding clusters indicate that this imagery event is flanked by two additional distinct stages. First, a stage prior to (inner) production (~300ms to ~400ms) featuring activity in both pre/motor and temporo-parietal regions. This is consistent with the inner loop of SFC theories, characterized by feedforward-feedback processes for speech planning[8,16,44]. And second, a stage following production (>500ms) featuring activity in posterior auditory as well as bilateral motor regions, consistent with a hypothesized feedback stage following motor execution[8,16,44]. As might be expected, the same analysis performed subtracting the *Reading* condition from the *Imagery* condition in source space removes much of the visual activity, but the exact same sequence can be observed (Fig S6). This again suggests that syllable decoding in the *Reading* condition was essentially driven by visual information, in contrast to the *Imagery* condition. Event 2 did not yield as clear a sequence (Fig S7 and Fig S8), as expected from the misalignments within and

across subjects and the weaker decoding results. It is nevertheless worth highlighting the implication of auditory, premotor, and motor areas as well as of area Sylvian parieto-temporal in this latter event, which suggests similar neural dynamics to event 1.

Zooming in on the inner loop for SFC

Our analyses revealed a sequence of neural representations leading to a speech imagery event consistent with SFC (Fig 2). As mentioned previously, a critical feature that distinguishes SFC from other influential models of speech production is the existence of an internal feedback loop characterized by the interplay between sensory and motor regions[8,16]. Although our previous analysis clearly shows activity in these regions prior to the expected time of inner production aligned with this, it represents but a snapshot of their dynamics. To examine these dynamics more closely, we extracted the time courses over event 1 of three key regions of interest (ROIs) in the left hemisphere, namely motor cortex, primary auditory cortex (core auditory), and posterior superior auditory cortex (posterior auditory) (*Methods*). Besides their hypothesized implication in the internal feedback loop, this reduced selection was motivated by the 'grounded' functions of these regions[53]. In other words, it is safe to attribute motor representations to motor cortical areas and auditory representations to auditory areas. Note that, in addition to a core auditory region, we selected a posterior auditory region for its known involvement in the computation of auditory feedback[54]. The anatomical location of the 3 ROIs was based on a well-known cortical atlas by Glasser et al.[55] (see *Methods* for the detailed procedure for the selection of the ROIs). In short, we selected, for each participant the MNI coordinates that displayed maximal activity within each of the corresponding atlas labels. Around each coordinate point, we then built a 4mm sphere and extracted the average time course of the sources within. We determined the times at which these ROIs' time courses were consistently activated across participants (significantly above their mean baseline activity; *Methods*) using a cluster-based permutation test (1000 permutations; one-tail). We expected the sequential activation of core auditory (coding for the acoustic representation), posterior auditory (coding for the prediction error), and motor areas (receiving feedback) to occur at least once before the expected time of imagery (i.e., during speech planning) and once after internal production.

Fig 5 shows the time courses of the selected ROIs (see Fig S9 for the activity of control regions and Fig S10 for additional auditory areas).

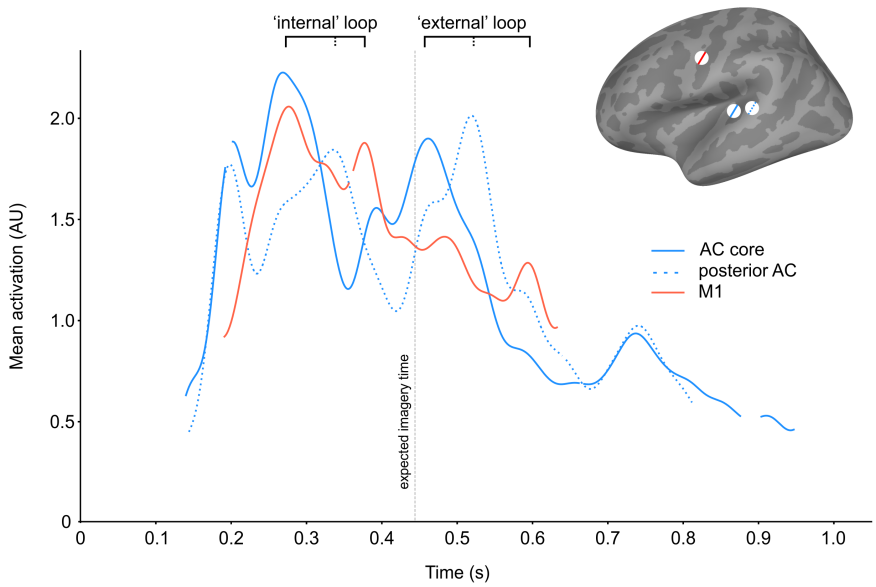

Fig 5. Time courses of auditory and motor regions during speech imagery. SFC theories hypothesize an internal feedback loop for motor planning and an external loop for post-production feedback, both characterized by feedforward and feedback processes between motor and auditory regions. The time courses of core auditory and motor regions in addition to a posterior

auditory region known for its role in sensory feedback show activity pre and post internal speech production in a sequence consistent with the hypothesized double feedback loop. Blank segments in the lines indicate non-significant times. Time courses were z-scored and low-pass filtered at 20 Hz (double pass Butterworth of order 5) for display purposes only.

The time courses show an initial peak of activity in auditory areas (~180ms) in line with the phonological encoding stage, followed by concurrent activity in both motor and core AC regions. This activity is consistent with the parallel activation of motor and auditory targets hypothesized by SFC models. This is followed by activity in posterior auditory regions and motor activation immediately before the expected time of imagery (vertical dashed line), consistent with an inner feedback stage prior to execution. Following motor activity, core auditory activity rises to peak at the expected time of imagery. We conjecture that this auditory activity (plus activity in secondary auditory regions; Fig S10) reflects the inner 'hearing' of the imagined syllable. The plot also shows that core auditory activity at the expected time of imagery is again closely followed by activity in the posterior auditory region and motor activity soon after that, consistent with the second feedback loop hypothesized by SFC and other models. In sum, the dynamics of auditory and motor regions are consistent with an 'internal' feedback loop prior to motor execution as well as an 'external' feedback loop following motor execution, both predicted by SFC theories[8,16,44]. The time courses of control ROIs (e.g., left visual cortex and frontal pole; Fig S9) show that these auditory and motor group-level dynamics are not a product of the analytical procedure.

A hierarchal organization of representations

Another distinguishing feature of recent SFC models for speech is the hierarchical organization of motor control units[8,16]. Two distinct tiers are hypothesized in accord with longstanding psycholinguistic models of speech production[56,57]. The higher tier codes for syllable-level speech information and features auditory targets. The lower tier features somatosensory targets and codes for information at the level of articulatory feature clusters roughly corresponding to phonemes. So far, our analyses showed a plausible sequence of neural representations leading to a speech imagery event in terms of number of processes, approximate occurrence times, and nature (e.g., whether they recur or evolve). Despite the many insights, however, the TG method remains relatively blind to the actual *length* of each representation because classifiers are trained on successive time points independently. Testing for the two distinct levels of representation hypothesized by SFC requires a method with greater sensitivity to the length of the representations involved during speech imagery.

To achieve this, we employed a similar multivariate pattern analysis approach (multinomial classification), but this time using sliding windows of different sizes (i.e., WMC). Specifically, we trained and tested a series of classifiers across the experimental trial using averages made from different number of data points. The rationale was that, if different representations respond to different lengths, better decoding at the times of these representations should be observed when the number of data points used for decoding (the window size) matches the length of the representation. Fig 6 shows the results of this approach using window sizes between 20ms and 300ms, averaged across participants (N = 21). Our choice of analysis windows was motivated, at the lower end, by knowledge of cortical transmission, estimated at a minimum of 20ms[58], and, at the upper end, by the length of the syllables produced by our participants, which rarely exceeded 300ms (0.5%).

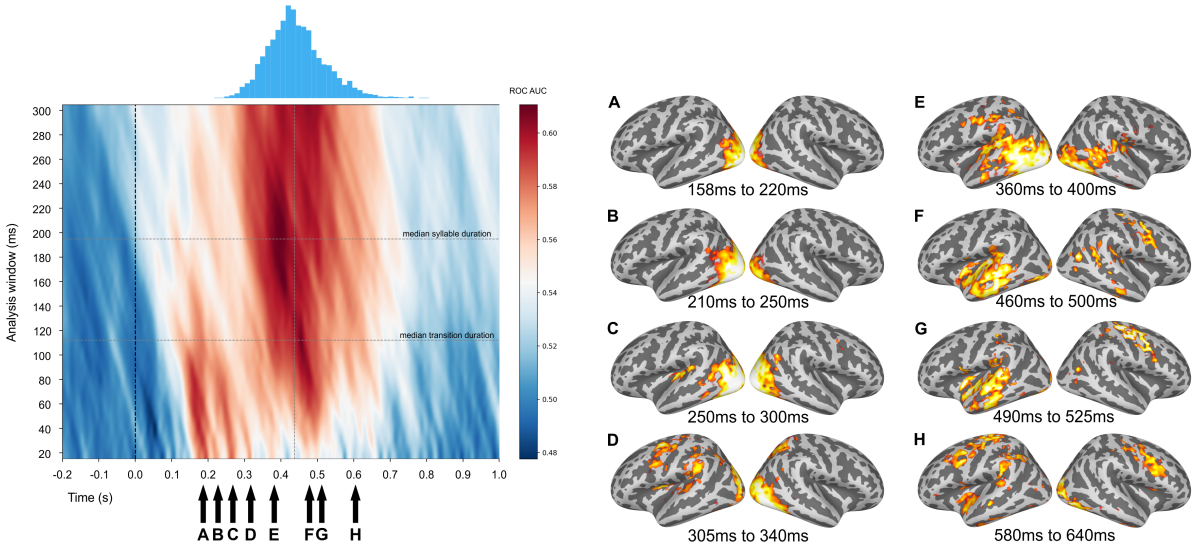

Fig 6. Representational similarity analysis reveals the optimal time and window length for decoding at each stage of the processing cascade. The matrix (left panel) shows the classifiers' ability to discriminate between a given syllable and the rest (e.g., *pa* from *ta* and *ka*) across event 1 (x-axis) when using averaged data of various lengths (*y*-axis). Analysis windows (*y*-axis) ranged between 20ms and 300ms with a step size of 10ms. Analysis times (x-axis) spanned between -0.2 and 1 sec, with a step size of 10ms. The analysis was performed within participant for each of the syllables using the same decoding scheme as before (cross-validation, regularization, classifier type, and scoring metric) except that each syllable was pit against the remaining two (e.g., *pa* vs *ta* and *ka*) rather than evaluating their pair-wise contrasts (*Methods*). The figure shows the grand average across participants. Results are plotted using spline interpolation between neighboring points for display purposes. The distribution of overt speech onsets is displayed on top of the matrix for reference as to the expected time of speech imagery (vertical dashed line at 436ms). The analysis highlights a sequence of distinct processing stages similar to Fig. 4 but with greater fidelity to time and additional insights into the length of each subroutine. While neural representations proceed rapidly in an encapsulated fashion leading up to the imagined event, as shown by processes A-E with optimal window sizes < 80ms, peak decoding is achieved around the expected speech imagery onset time (436ms; vertical dashed line) with a window length of 200ms, which closely matches the median syllable duration across the cohort (horizontal dashed line at 194ms; interquartile range: 59ms; Table S1). A second peak decoding cluster near the expected imagery time at window size of 110ms indicates a predominance of the syllables' consonant–vowel transitions in the imagery representation (horizontal dashed line at 114ms indicates participants' median transition duration; see main text for further details). The progression of source-projected evoked responses (right panel, A-H) faithfully reproduces the previously identified (Fig 4) with the addition of a right-lateralized inferior frontal and motor activity (G-H) following the auditory speech imagery event (F-G), in line with Tourville et al.[59] and current models of speech production[46].

As before, the analysis highlights the rapid succession of representations leading to the speech imagery event. In particular, the earlier set of processes (marked A – E, Fig 6), corresponding to the phonological, speech planning, internal loop iterations, and motor execution stages, appear to be relatively brief and self-contained, with successive onsets every 40ms-60ms and short-lived representations (optimal decoding windows under ~80ms). The fact that decoding accuracy does not increase when analysis windows span neighboring processes again suggests that the representations involved are highly specific (are not additive). More interestingly, peak decoding accuracy in this analysis was achieved close to expected time for imagery (~390ms) with an optimal decoding window of 200ms, a value which closely matches the median syllable length across participants in their overt productions (194ms; interquartile range: 59ms). This suggests that, while processes leading up to the speech imagery event involve fleeting and evolving representations, the speech imagery event itself unfolds at the natural rate of the speaker. Somewhat conversely, this implies that speech planning entails a compressed form of the utterance to be produced, even at the level of a single syllable. The fact that optimal decoding was achieved at syllable length even though the syllables used were matched in their vowels (i.e., in their last phonemes: *pa*, *ta*, and *ka*) also indicates that what was presumably the same vowel (vowel *a* in all cases) may be realized differently in 'the mind's ear' after each of the consonants (*p*, *t*, and *k*) as it happens during overt production. On the other hand, the runner-up decoding cluster, also located near the expected time of imagery (~450ms), was related to a window-size of 110ms, which closely matches the median consonant-vowel transition length (114ms), as estimated from participants'

overt productions (Table S1). This indicates that, although the vowel may also be represented at the expected time of imagery, the transitions still dominate that representational space, as could be predicted from the set of syllables employed.

If our last conjectures are correct, we should expect syllables varying in their vowels while keeping their consonants constant (e.g., *ta*, *tu*, *ti*) to behave rather differently. Specifically, we should see that, although similar in their planning stage (compressed representations), the optimal window for decoding of these new syllables at the expected imagery time is larger, reflecting the more dilated nature of vowels in acoustic space compared to consonants. Additionally, the cluster at the consonant-vowel transition level should be reduced and the optimal time for decoding should be delayed in respect to the expected imagery time, both reflecting a greater weight on the vowel (i.e., latter part of the syllable) rather than on the transition.

To test these predictions, we collected data from a new cohort of participants (N = 9; 7 women; mean age = 23; *std* = 7.94) imagining syllables varying in their vowels rather than in their consonants (specifically, *ta*, *tu*, and *ti*). Importantly, we first replicated the TG results on this new data set (Fig S12; cf. Fig 3). We then run the same WMC analysis (Fig 7).

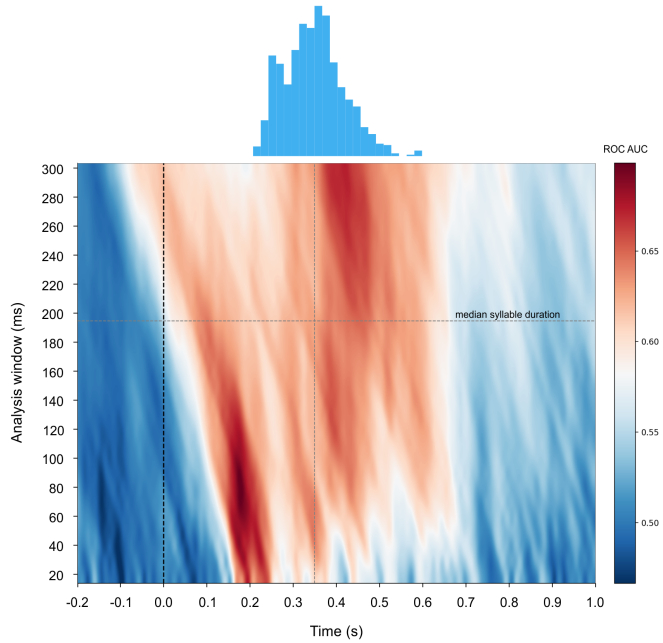

Fig 7. Average syllable discriminability over time for syllables *ta*, *tu*, and *ti*. The matrix shows the classifiers' ability to discriminate between a given syllable and the rest (e.g., *ta* from *tu* and *ti*) across event 1 (x-axis) when using averaged data of various lengths (*y*-axis). Analysis windows (*y*-axis) ranged between 20ms and 300ms with a step size of 10ms. Analysis times (x-axis) spanned between -0.2 and 1 sec, with a step size of 10ms. The figure shows the grand average across participants. Results are plotted using spline interpolation between neighboring points for display purposes. The distribution of overt syllable onsets is displayed on top of the matrix for reference as to the expected time of inner speech production. In contrast to the analogous analysis for syllables *pa*, *ta*, and *ka*, higher decoding accuracies around phonological encoding (~200ms) and after the expected time of imagery (vertical dashed line at 349ms) over large analysis windows (> 240ms) suggest a greater weight on acoustic rather than motor representations.

In a context of overall higher syllable decoding accuracy (Fig S13), the strongest cluster was found in the early stages of speech planning, that is in times aligned with phonological encoding[51], reflecting the greater representation of the syllables in acoustic space. In line with this, the cluster at the transition level, despite occurring at the expected time (~400ms), was also greatly minimized. Moreover, as predicted, the strongest decoding was delayed in respect to the expected imagery time (349ms; vertical dashed line, Fig 7) highlighting the greater weight of the vowel in this syllable set. Interestingly, higher decoding accuracy around the expected imagery time was achieved when using relatively long window sizes (>240ms) compared to the median syllable duration (195ms; *iqr* = 43ms). This may reflect the

arbitrary lengths of the vowels imagined by participants, which is also reflected in the protracted temporal generalization scores during the actual imagined event (Fig S12). Finally, if acoustic space carries greater weight in the imagery representations for this latter set of syllables (*ta*, *tu*, *ti*), we should observe that the most discriminable syllable from the set is that which is most distinct in that representational space (i.e., *ti*; Fig S14). The WMC analysis separated by syllable (Fig S15) indicates that this is the case. Highest and most sustained discriminability (indicated by the vertical distance between the lines) was observed for the syllable *ti*, followed by *ta* and lastly *tu*. The fact that the syllable *ka*, whose consonant *k* is most similar in place of articulation to the vowel *i* of *ti* (velar constriction vs raised posterior tongue) as its discriminating factor from the remaining two syllables (*pa* and *ta*), was the least discriminable of the *pa-ta-ka* set (Fig S16) again supports the idea that acoustic rather than articulatory space governed the decoding of the *ta-tu-ti* set. This contrasts with the *pa-ta-ka* set, where data suggests a greater balance between acoustic and motor representational spaces. In all, our results strongly support the hypothesis of two tiers of representation during speech imagery in line with recent SFC models.

Discussion

Speech imagery refers to the capacity to internally 'hear' self-generated speech. Despite its pervasive nature in many aspects of cognition (e.g., predictive processing), the study and use of speech imagery for clinical or academic research has been tremendously challenging. In this work, we sketch out the dynamics of speech imagery by pairing MEG with a novel experimental protocol designed to overcome known methodological difficulties. By decoding participants' imagined utterances, we show that producing imagined speech involves a rapid succession of relatively encapsulated neural representations that neatly maps onto the predictions of current theoretical models of overt production.

As a snapshot, the neural correlates we report for these fleeting representations are consistent with previous fMRI research on speech imagery, inner, and covert speech[7,10,60]. An advantage of our time-resolved approach is that it can additionally provide a dynamic picture of speech imagery. In broad strokes, our data is consistent with two well-defined stages during imagery, namely planning and (internal) production (Fig 4). The production stage is characterized by widespread left-lateralized auditory activity at the expected time of imagery (~400ms, based on participants' overt productions of the same syllables) immediately preceded by speech motor activity. We conjecture this auditory activity to correspond to the percept associated with speech imagery[11,19]. Following the imagined event, our syllable decoding approach identified two additional time periods with distinct neural representations, associated with pSTG activity (~500ms) and subsequent bilateral pre/motor activity (~570ms). This pattern of activity is consistent with a hypothesized external feedback loop during overt production, in which corrective error from comparing the predicted sensory consequences of planned articulatory gestures with their actual consequences (i.e., auditory feedback) is forwarded to motor regions[8,44,46]. Indeed, the location of the posterior auditory cluster is consistent with fMRI research using altered feedback to identify error-related activity[59,61] as well as recent ECoG research that distinguishes sensory processing (in more anterior regions) from feedback error signals[62]. Activity in bilateral pre/motor regions has also been reported following auditory error[59,63], in line with our observations. Interestingly, while error-related activity can be modulated experimentally by altering feedback in both overt[59,63] and imagined speech[13], it is unclear what the expectation should be in the case of imagery when no matching or mismatching information is given. In accord with a recent hypothesis on predictive processing[21], our intuition was that, since there is no overt auditory feedback to meet predictions (i.e., there is less input than predicted), an error response in auditory regions should still be produced. Our results support this hypothesis, highlighting the potential of our approach for research on predictive processing.

In terms of (internal) speech planning, we identified at least three distinct time periods of significant syllable decoding prior to the production stage (Fig 4). The first, ~180ms after syllable presentation, was associated with activity in left posterior temporal cortex and thus consistent, both in time and location, with a much-theorized phonological encoding stage[16,50,51]. Phonological encoding may thus be present whether production is internally generated (e.g., from abstract thought) or externally triggered (e.g., from reading). The second was characterized by concurrent activity in left auditory and bilateral

motor regions at ~260ms. Such activity is predicted by SFC models recently applied to speech motor control in which auditory and motor targets are accessed in parallel immediately following phonological encoding[16]. SFC also predicts an internal feedback loop prior to motor execution, in which sensory predictions (auditory and somatosensory) are compared to the intended sensory targets[16,44,64]. Although theoretically well-grounded, there is yet no direct empirical support for an internal loop in speech production. Indirect evidence comes from the timings with which individuals correct themselves during speech errors, which are too fast for responses to external auditory feedback[65]. Activity at ~300ms (Fig 4) is consistent with the hypothesized internal feedback loop, featuring the posterior auditory cluster, the supramarginal gyrus, and pre/motor regions. This result is further supported when examining the time courses of auditory and motor areas during imagery (Fig 5). Specifically, the sequence of activations observed for the putative external loop is also present immediately before the expected time of imagery, consistent with the monitoring and planning role of the internal loop.

In all, the generation of speech imagery appears to closely mirror that of overt speech, with full-blown planning and (internal) execution stages. Our data are thus inconsistent with views of imagery as a by-product of motor planning (e.g., [49]). Moreover, given the presence of unspecific micromovements at the expected time of imagery (Fig S2-S4), our hypothesis is that, during imagery, speech plans are executed but aborted (inhibited) at the periphery. In this sense, speech imagery may be seen as analogous to concrete (as opposed to abstract) forms of inner speech (cf., [7]).

Because of the neuroimaging method (i.e., MEG), it may be necessary to question the accuracy of the reported cortical areas. Although possible, inaccuracies seem unlikely given the close correspondence between the main sources of activity in our imagery task and previously reported clusters in motor and sensory areas[60]. This is particularly so in the case of auditory regions, which also show a unified pattern of activity and consistency with the expected time of imagery (Fig S10). A possible exception could be the posterior auditory cluster because of its proximity to area Spt (cf., [66]). Indeed, the behavior of this cluster could be seen as consistent with Spt's hypothesized role in auditory-motor transformations ([8,67]). We also found activity in premotor and motor regions to be in very close spatial proximity (Fig 4). It is therefore possible that the activity we attribute to the motor cluster pertains to premotor cortex instead. This would be consistent with a hypothesized origin of sensory predictions in the premotor cortex in response to inputs (efference copies) from motor areas as well as with known inputs of premotor areas to motor cortical regions for production[46].

Using a novel WMC approach, two different but concurrent optimal decoding windows were found at the expected time of imagery, corresponding to the length of overtly produced syllables and consonant-vowel transitions (~200ms and ~100ms, respectively, estimated from our data) (Fig 6 and Fig 7). These results are thus consistent with nested levels of motor control so far only hypothesized by hierarchical SFC models of speech production[8,16] but more generally found in the motor literature (e.g., hand gestures[47,68,69]). In line with speech models, our analysis of the *ta-tu-ti* set also suggests that syllable-level information is acoustically represented. This implies a level of speech motor control driven by auditory representations at the syllable level that aligns well with the observed auditory and motor activity. A tentative possibility is that activity in ventral somatosensory and supramarginal regions (Fig S11), which closely mimics that of core and posterior auditory areas (respectively), reflects the lower (i.e., somatosensory) level of motor control[16]. By this account, previously reported inferior parietal activity prior to speech imagery events (e.g., [11,15]) could relate to this level of control.

Another interesting finding revealed by the WMC analysis is the compressed nature of speech representations during planning, which contrasts with the natural rate at which the internal production unfolds (cf., [70]). There is evidence that inner speech and speech imagery encode tempo, pitch, timber, and loudness information[13,14,38,71] but little is known about the relationship between production and planning stages in imagined, inner, or overt speech. Both sets of syllables (*ta-tu-ti* and *pa-ta-ka*) exhibited a similar pattern of optimal decoding windows under 100ms during planning and ≥200ms during execution suggesting compressed representations during planning at least for short (syllable-length) utterances. This is important for potential brain-computer interfaces, which could account for this feature to increase decoding performance.

Finally, some methodological considerations. When it comes to decoding imagined speech, we would like to emphasize the importance of aligning responses to reduce the amount of noise in the data with which classifiers are trained. We took several measures by design to improve the alignment of imagined events both within and between participants (*Methods*) and sustain that many of the difficulties in decoding imagery from broadband signals are due to temporal misalignments. Indeed, many speech decoding studies using time-resolved methods (e.g., EEG, MEG) have ultimately resorted to frequency analyses (e.g., power modulations, cross-frequency couplings)[72], loosing temporal resolution. While our objective was not decoding per se, we were able to decode participants' imagined utterances with relatively high accuracies from non-invasive data using simple linear classifiers. Although we feel excited by recent invasive approaches and sophisticated analysis pipelines (e.g., [40]), we would like to sound a note of hope for non-invasive methods and 'lighter' analytical procedures. Another important challenge in decoding imagined (or planned) speech is determining the optimal temporal window for analysis. Here, we took a primary focus on the syllable as the basic unit for decoding, although our data turned out to be consistent with at least two representational levels. Motor-based representations in the order of consonant-vowel transitions feature in recent ECoG research[73]. However, a focus on syllable-like chunks may have several advantages worth considering. In contrast to smaller or larger units, syllables offer a remarkable rhythmicity (around 4-5 Hz) across languages[74–76], a feature that could be leveraged for the decoding of continuous imagined speech particularly with the incorporation of language models to constrain decoders' options (e.g., [33,40]). Although potentially more numerous than smaller units, most common utterances can be produced by the combination of a reduced number of syllables[77]. Syllables are also relatively stable and less variable than smaller units[78,79]. Finally, while the existence of a motor 'syllabary' is still an empirical question, our data suggests that syllables are represented in multiple spaces concurrently (at least motor and auditory). This characteristic makes them not only more robust to degradation than kinematic representations (e.g., in amyotrophic lateral sclerosis) but also more likely to be decoded by imaging methods that provide wide brain coverage (e.g., EEG, MEG).

In all, our results show an evolving sequence of representations for speech imagery subserved by neural dynamics akin to SFC. It is assumed that the same sequence underlies the generation of sensory predictions through the speech motor system that modulates speech perception and subserves the articulatory loop of phonological working memory. Our results thus highlight the potential of speech imagery for research, granted the appropriate experimental approaches and analytical methods, and paves the way for successful clinical and industrial applications.

1. Aristotle & Lawson-Tancred, H. (Trans. . *De Anima (on the soul)*. (Penguin Books, 1986).
2. Hobbes, T. *Leviathan*. (Penguin Books).
3. Hume, D. A treatise of human nature. in (ed. Mossner, E. C.) (Penguin Books, 1969).
4. James, W. *The principles of psychology*. (MacMillan, 1890).
5. Wundt, W. M. *Elemente der Völkerpsychologie: Grundlinien einer psychologischen Entwicklungsgeschichte der Menschheit.* (Kröner, 1913).
6. Fernyhough, C. & Alderson-Day, B. Inner Speech: Development, Cognitive Functions, Phenomenology, and Neurobiology. *Psychol. Bull.* **141**, 931–965 (2015).
7. Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M. & Lœvenbruck, H. What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behav. Brain Res.* **261**, 220–239 (2014).
8. Hickok, G., Houde, J. & Rong, F. Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron* **68**, 1–28 (2011).
9. Hickok, G. The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model. *J. Commun. Disord.* **45**, 393–402 (2012).
10. Price, C. J. NeuroImage A review and synthesis of the fi rst 20 years of PET and fMRI studies of heard speech , spoken language and reading. *Neuroimage* **62**, 816–847 (2012).
11. Tian, X. & Poeppel, D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* **1**, 1–23 (2010).
12. Tian, X. & Poeppel, D. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Front. Hum. Neurosci.* **6**, 314 (2012).
13. Tian, X. & Poeppel, D. Dynamics of Self-monitoring and Error Detection in Speech Production: Evidence from Mental Imagery and MEG. *J. Cogn. Neurosci.* **27**, 352–364 (2015).
14. Tian, X., Ding, N., Teng, X., Bai, F. & Poeppel, D. Imagined speech influences perceived loudness of sound. *Nat. Hum. Behav.* **2**, 225–234 (2018).
15. Tian, X., Zarate, J. M. & Poeppel, D. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex* **77**, 1–12 (2016).
16. Hickok, G. Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* **13**, (2012).
17. Blom, T., Feuerriegel, D., Johnson, P., Bode, S. & Hogendoorn, H. Predictions drive neural representations of visual events ahead of incoming sensory information. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 7510–7515 (2020).
18. Kilteni, K., Andersson, B. J., Houborg, C. & Ehrsson, H. H. Motor imagery involves predicting the sensory consequences of the imagined movement. *Nat. Commun.* **9**, 1–9 (2018).
19. Grush, R. The emulation theory of representation: Motor control, imagery, and perception. *Behav. Brain Sci.* **27**, 377–396 (2004).
20. Moulton, S. T. & Kosslyn, S. M. Imagining predictions : mental imagery as mental emulation Email alerting service Imagining predictions : mental imagery as mental emulation. *Philos. Trans. R. Soceity B* **364**, 1273–1280 (2009).
21. Keller, G. B. & Mrsic-Flogel, T. D. Predictive Processing: A Canonical Cortical Computation. *Neuron* **100**, 424–435 (2018).
22. Friston, K. A theory of cortical responses. *Philos. Trans. R. Soceity B* **360**, 815–836 (2005).
23. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
24. Wolpert, D. M. & Ghahramani, Z. Computational principles of movement neuroscience. *Nat. Neurosci.* **3**, 1–6 (2000).
25. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science (80-.).* **275**, 1593–1600 (1997).
26. Montague, P. R., Dayan, P. & Sejnowski, J. A Framework for Mesencephalic Predictive Hebbian Learning. *J. Neurosci.* **76**, 1936–1947 (1996).
27. Niv, Y. & Montague, P. R. Theoretical and Empirical Studies of Learning. 329–350 (2008).
28. Orpella, J. *et al.* Integrating when and what information in the left parietal lobule allows language rule generalization. *PLoS Biol.* **18**, 1–26 (2020).
29. Orpella, J., Mas-Herrero, E., Ripollés, P., Marco-Pallarés, J. & de Diego-Balaguer, R. Language statistical learning responds to reinforcement learning principles rooted in the striatum. *PLoS Biol.* **19**, 1–23 (2021).
30. Jones, S. R. & Fernyhough, C. Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Conscious. Cogn.* **16**, 391–399 (2007).
31. Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
32. Makin, J. G., Moses, D. A. & Chang, E. F. With an Encoder – Decoder Framework. *Nat. Neurosci.* **23**, (2020).
33. Moses, D. A., Leonard, M. K., Makin, J. G. & Chang, E. F. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nat. Commun.* **10**, (2019).
34. Brumberg, J. S. & Guenther, F. H. Development of speech prostheses: current status and recent advances. *Expert Rev Med Devices* **7**, 667–679 (2010).
35. Guenther, F. H. *et al.* A wireless brain-machine interface for real-time speech synthesis. *PLoS One* **4**, (2009).
36. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
37. Martin, S., Iturrate, I., Millán, J. del R., Knight, R. T. & Pasley, B. N. Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Front. Neurosci.* **12**, 1–10 (2018).
38. Tian, X. & Poeppel, D. The effect of imagination on stimulation: the functional specificity of efference copies in

speech processing. *J. Cogn. Neurosci.* **25**, 1020–1036 (2013).
39. Jack, B. N. *et al.* Inner speech is accompanied by a temporally-precise and content-specific corollary discharge. *Neuroimage* **198**, 170–180 (2019).
40. Moses, D. A. *et al.* Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *N. Engl. J. Med.* **385**, 217–227 (2021).
41. King, J. R. & Dehaene, S. Characterizing the dynamics of mental representations: The temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).
42. Fyshe, A. Studying language in context using the temporal generalization method. *Philos. Trans. R. Soceity B* **375**, (2019).
43. Poeppel, D. & Assaneo, M. F. Speech rhythms and their neural foundations. *Nat. Rev. Neurosci.* **21**, (2020).
44. Houde, J. F. & Nagarajan, S. S. Speech production as state feedback control. *Front. Hum. Neurosci.* **5**, 1–14 (2011).
45. Wolpert, D. M., Ghahramani, Z. & Jordan, M. I. An internal model for sensorimotor integration. *Science (80-.).* **269**, 1880–1882 (1995).
46. Guenther, F. H. *Neural Control of Speech*. *Neural Control of Speech* (MIT Press, 2016). doi:10.7551/mitpress/10471.001.0001
47. Grafton, S. T. The cognitive neuroscience of prehension: Recent developments. *Exp. Brain Res.* **204**, 475–491 (2010).
48. Nikulin, V. V., Hohlefeld, F. U., Jacobs, A. M. & Curio, G. Quasi-movements: A novel motor-cognitive phenomenon. *Neuropsychologia* **46**, 727–742 (2008).
49. Jeannerod, M. Mental imagery in the motor context. Special Issue: The neuropsychology of mental imagery. *Neuropsychologia* **33**, 1419–1432 (1995).
50. Hickok, G. & Poeppel, D. The cortical organization of speech processing. *Nat. Rev. Neurosci.* **8**, 393–402 (2007).
51. Indefrey, P. & Levelt, W. J. M. The spatial and temporal signatures of word production components. *Cognition* **92**, 101–144 (2004).
52. Hickok, G., Buchsbaum, B., Humphries, C. & Muftuler, T. Auditory-motor interaction revealed by fMRI: Speech, music, and working memory in area Spt. *J. Cogn. Neurosci.* **15**, 673–682 (2003).
53. Harnad, S. The symbol grounding problem. *Phys. D* **42**, 335–346 (1990).
54. Forseth, K. J., Hickok, G., Rollo, P. S. & Tandon, N. Language prediction mechanisms in human auditory cortex. *Nat. Commun.* **11**, 1–14 (2020).
55. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
56. Levelt, W. J. M. Spoken word production: A theory of lexical access. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 13464–13471 (2001).
57. Dell, G. S. A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychol. Rev.* **93**, 283–321 (1986).
58. Buzsáki, G. Neural syntax: cell assemblies, synapsembles and readers. *Neuron* **68**, (2010).
59. Tourville, Reilly & Guenther. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage* **39**, 1429–1443 (2008).
60. McNorgan, C. A meta-analytic review of multisensory imagery identifies the neural correlates of modality-specific and modality-general imagery. *Front. Hum. Neurosci.* **6**, 1–14 (2012).
61. Behroozmand, R. *et al.* Sensory-motor networks involved in speech production and motor control: An fMRI study. *Neuroimage* **109**, 418–428 (2015).
62. Forseth, K. J., Hickok, G., Rollo, P. S. & Tandon, N. Language prediction mechanisms in human auditory cortex. *Nat. Commun.* **11**, 1–14 (2020).
63. Ozker, M., Doyle, W., Devinsky, O. & Flinker, A. A cortical network processes auditory error signals during human speech production to maintain fluency. *PLOS Biol.* **20**, e3001493 (2022).
64. Hickok, G., Houde, J. & Rong, F. Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron* **69**, 407–422 (2011).
65. Nozari, N., Dell, G. S. & Schwartz, M. Is comprehension necessary for error detection? A conflict- based account of monitoring in speech production. *Cogn. Psychol.* **63**, 1–33 (2011).
66. Buchsbaum, B. R. *et al.* Conduction aphasia, sensory-motor integration, and phonological short-term memory - An aggregate analysis of lesion and fMRI data. *Brain Lang.* **119**, 119–128 (2011).
67. Hickok, G., Okada, K. & Serences, J. T. Area Spt in the human planum temporale supports sensory-motor integration for speech processing. *J. Neurophysiol.* **101**, 2725–2732 (2009).
68. Klatzky, R. L., McCloskey, B., Doherty, S., Pellegrino, J. & Smith, T. Knowledge about hand shaping and knowledge about objects. *J. Mot. Behav.* **19**, 187–231 (1987).
69. Rizzolatti, G. *et al.* Functional organization of inferior area 6 in the macaque monkey - II. Area F5 and the control of distal movements. *Exp. Brain Res.* **71**, 491–507 (1988).
70. Lashley K.S. The problem of serial order in behavior. *Cereb. Mech. Behav.* 112–136 (1951).
71. MacKay, D. G. Constraints on theories of inner speech. in *Auditory imagery.* 121–149 (Lawrence Erlbaum Associates, Inc, 1992).
72. Proix, T. *et al.* Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nat. Commun.* **13**, 1–14 (2022).
73. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
74. Ding, N. *et al.* Temporal modulations in speech and music. *Neurosci. Biobehav. Rev.* **81**, 181–187 (2017).
75. Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J. & Lorenzi, C. A cross-linguistic study of speech modulation spectra. *J. Acoust. Soc. Am.* **142**, 1976–1989 (2017).
76. Greenberg, S., Carvey, H., Hitchcock, L. & Chang, S. Temporal properties of spontaneous speech - A syllable-

centric perspective. *J. Phon.* **31**, 465–485 (2003).
77. Schiller, N. O., Meyer, A. S., Baayen, R. H. & Levelt, W. J. M. A comparison of lexeme and speech syllables in Dutch. *J. Quant. Linguist.* **3**, 8–28 (1996).
78. Browman, C. P. & Goldstein, L. Some notes on syllable structure in articulatory phonology. *Phonetica* **45**, 140–155 (1988).
79. Byrd, D. C-Centers Revisited. 285–306 (1995).