

OPEN

Motor representations underlie the reading of unfamiliar letter combinations

Alan Taitz^{1*}, M. Florencia Assaneo^{2,4}, Diego E. Shalom^{3,1} & Marcos A. Trevisan^{1,3}

Silent reading is a cognitive operation that produces verbal content with no vocal output. One relevant question is the extent to which this verbal content is processed as overt speech in the brain. To address this, we acquired sound, eye trajectories and lips' dynamics during the reading of consonant-consonant-vowel (CCV) combinations which are infrequent in the language. We found that the duration of the first fixations on the CCVs during silent reading correlate with the duration of the transitions between consonants when the CCVs are actually uttered. With the aid of an articulatory model of the vocal system, we show that transitions measure the articulatory effort required to produce the CCVs. This means that first fixations during silent reading are lengthened when the CCVs require a greater laryngeal and/or articulatory effort to be pronounced. Our results support that a speech motor code is used for the recognition of infrequent text strings during silent reading.

The faculty of language entails a repertoire of mental operations known as inner speech, in which verbal content is produced but voice is inhibited¹. This internal production of words is ubiquitous: “as you read this text, the chances are you can hear your own inner voice narrating the words. You may hear your inner voice again when [...] imagining how a phone conversation this afternoon will play out”². One relevant question regarding speech-related tasks is the extent to which they are processed as actual speech in the brain. During speech perception, for example, neural patterns are organized around acoustic features and do not contain articulatory representations as the ones produced during speech³. On the contrary, speech imagery has been recently associated with the production of efference copies², a specific signature of motor patterns. Moreover, spectrotemporal features of inner speech were decoded with significant predictive accuracy from models built from overt speech data⁴.

Increasing evidence suggests that, despite the absence of vocal articulation, motor patterns form part of inner speech. However, investigation on this matter has remained challenging, due in part to the lack of behavioral output of speech imagery, which makes difficult to time-lock precise events (acoustic features, phonemes, words) to neural activity⁴. Silent reading offers a direct way to tackle this problem, since it provides us with the trajectory of the reader's eyes along the text as a natural behavioral output⁵. A fair amount of experiments showed that phonology affects silent reading^{6,7} and, beyond the behavioral level, it has been also shown that visual recognition of words activates sub-phonemic features at brain level⁸. Eye movements in oral and silent reading have also been investigated⁹, offering a unique opportunity to understand the specific processing features of each operation. Here we capitalize on this background to advance a quantitative study on the relation between articulatory and ocular variables during reading, which is still lacking.

We hypothesize that articulatory simulations underlie the decoding of unfamiliar letter strings during silent reading. We tested this by a quantitative exploration of ocular and articulatory dynamics during oral and silent reading of consonant-consonant-vowel (CCV) trigrams.

Ocular trajectories consist of gaze fixations separated by rapid eye movements called saccades¹⁰. Here we characterized reading dynamics by computing standard timing variables such as the duration of first fixations and the integration of the successive fixations on each CCV, which are standard measures that have been largely used to disentangle visual, lexical and contextual processing in reading^{11,12}.

¹Physics Institute of Buenos Aires (IFIBA) CONICET, Buenos Aires, Argentina. ²Department of Psychology, New York University, New York, NY, 10003, USA. ³Department of Physics, University of Buenos Aires (UBA), Buenos Aires, 1428EGA, Argentina. ⁴Instituto de Neurobiología, UNAM, Campus Juriquilla, Querétaro, México. *email: taitz@df.uba.ar

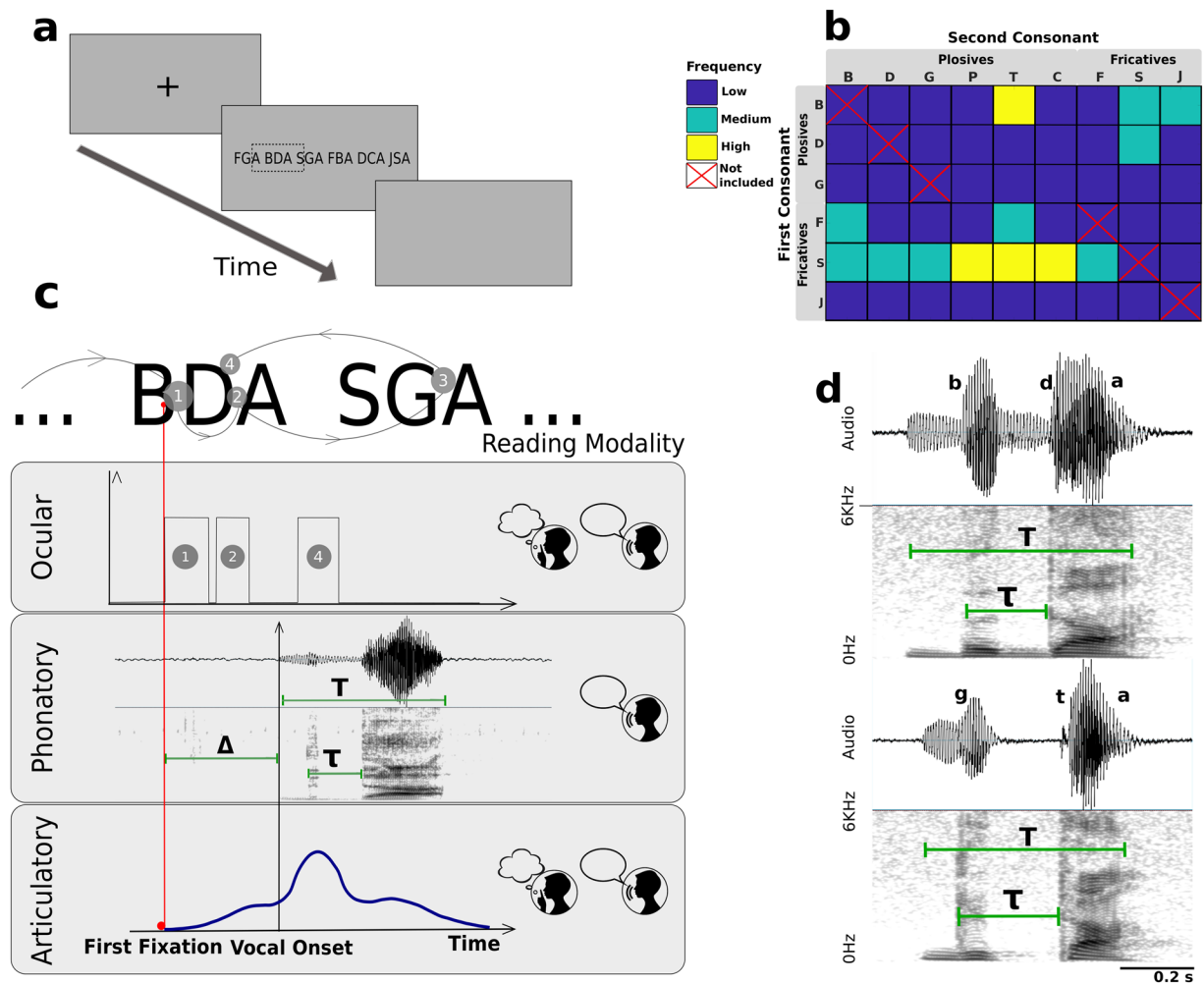


Figure 1. Monitoring ocular, phonatory and articulatory variables during reading. **(a)** After fixing the sight on a cross in the middle of the screen, a sequence of 6 CCVs appeared for the participant to read. The operation was repeated until three repetitions of the complete CCV set were presented. In one block, participants read orally and in the other block they read silently. **(b)** Intra-word frequency f of the CCVs in Spanish (in appearances per million words). The range was discretized in three levels: low ($f < 10$), medium ($10 < f < 50$) and high ($f > 50$). **(c)** Three types of variables were measured during reading. Ocular: duration of first fixation (FFD = 1); duration of fixations on a CCV prior to fixating a following one (FPRT = 1 + 2); and total fixation duration (TFT = 1 + 2 + 4). Phonatory: We measured the delay between the first fixation and the vocal onset Δ , the transition time between consonants τ , and the total pronunciation time T (oral reading). Articulatory: Lip movement was registered by an accelerometer fixed to the lower lip. **(d)** Spectrograms and phonatory variables of *bda* (voiced plosives) and *gta* (unvoiced second consonant).

To emphasize articulatory dynamics, focus was put on CCVs formed with plosive consonants, which require sharp movements to close the vocal tract completely. For instance, when the lips occlude the air passage, consonant *b* is produced; in this case, also the vocal folds are vibrating. Instead, when folds are not vibrating, the same occlusion produces the unvoiced plosive *p*¹³. Plosives therefore involve sharp articulatory movements combining on-off folds oscillations¹⁴, and sequences of plosives and consonants have also been modeled and synthesized from physical principles^{15,16}. From a lexical point of view, plosive CCVs form a homogeneous subset of pseudo-words with low intra-word frequency values¹⁷. This combination of motor, mathematical and lexical features makes these CCVs an ideal set of stimuli to explore the relationship between silent reading and vocal articulation.

Results

Measured Variables. Thirty native Spanish speakers read a set of consonant-consonant-vowel structures (CCV) from a computer screen. The set was formed by every combination of fricative consonants (*s*, *f*, *j*) and plosive consonants (*b*, *d*, *g*, *p*, *t*, *k*) ending with the vowel *a*. A few examples of these trigrams are *fga* (fricative-plosive), *bda* (plosive-plosive) and *jsa* (fricative-fricative), shown in Fig. 1a. None of the trigrams is a Spanish word (intra-word frequencies are displayed in Fig. 1b).

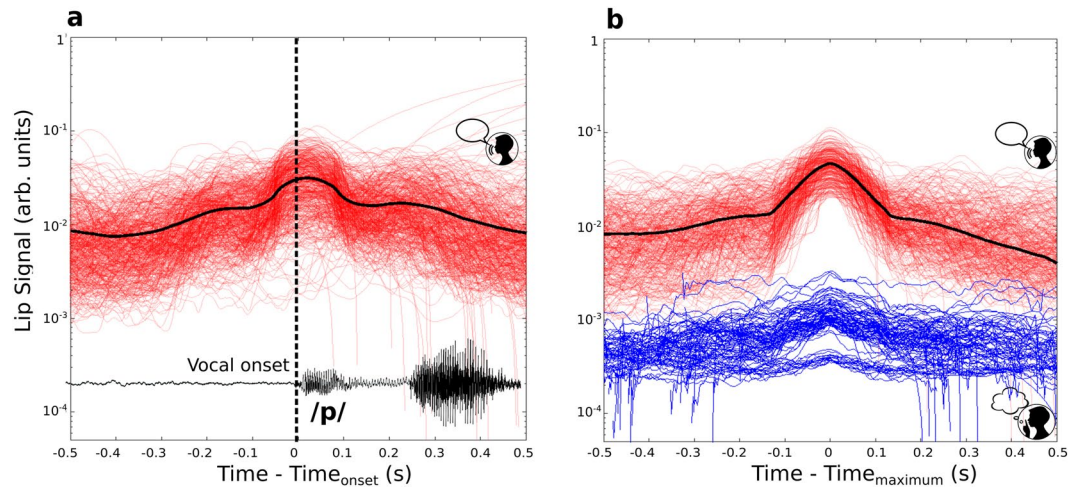


Figure 2. Lip movement is strongly inhibited during silent reading. **(a)** Red traces are the accelerometer signals during oral reading, aligned at vocal onset (absolute value, smoothed by a 140 ms moving window). Single trials are represented in red and averaged in black. **(b)** In order to compare ranges of articulation of the different reading modalities, the signals were realigned to the maximum. Oral reading is displayed in red and its average in black while silent reading in blue. At maxima, traces exhibit a difference of roughly two orders of magnitude.

The experiment was divided into two blocks, each formed with three repetitions of the complete set of CCVs. In one block the participants read the set aloud, and in the other they were instructed to read in silence. Blocks and CCVs were randomized before presentation. For each CCV we measured the following variables, sketched in Fig. 1c (videos of the experiment are available at Supplementary Videos S1 to S5):

Ocular (oral and silent blocks). Eye movements were recorded and the following variables were computed for each CCV: 1. the duration of the first fixation (FFD); 2. the duration of all fixations prior to passing to another CCV (First Pass Reading Time, FPRT) and 3. the total fixation time (TFT). These are standard ocular variables used to characterize cognitive processing in reading; the first two are typically considered early measures, whereas measure 3 reflects later processing stages^{18,19}.

Phonatory (oral block). From the spectrograms of the audio records, we computed: 1. the delay Δ between the onset of visual fixation and voice onset; 2. the total duration T of the CCV, and 3. a consonantal transition τ , shown in detail in Fig. 1d for the plosive-plosive combinations *bda* and *gta*. We defined τ as the interval between the release of the first occlusion, characterized by a brief voiced sound, and the release of the second one, that give rise to the vowel *a*. The transition τ does not represent the total duration of the vocal tract movements, but instead a portion of the dynamics between consonants that can be reliably extracted from the sound spectra. Finally, differences in speech rate across subjects were washed out by normalizing the transitions to the total duration of the CCV, $\tau' = \tau/T$, leaving us with the pre-phonatory variable Δ and the phonatory variable τ' .

Articulatory (oral and silent blocks). Lip movement was acquired by attaching an accelerometer to the lower lips with medical tape.

Articulation is inhibited during silent reading. We first investigated the articulatory activity in both oral and silent reading blocks. The main vocal tract articulators are the lips, the jaw and the tongue. Lip movements can be obtained from the exterior of the mouth, minimizing the interference of the measuring device with articulatory dynamics. Since articulators are coordinated during speech²⁰, we assumed that the range of lip movement is a good estimate for the range of the whole articulatory movements.

In Fig. 2 we show the dynamics of the lips for the CCVs that start with the bilabial consonant *p* (*pba*, *pda*, *pga*, *pta* and *pca*), which require maximal lip displacement. In Fig. 2a we show the absolute value of the accelerometer signals during oral reading, aligned to the beginning of the vocalization. We recovered the characteristic dynamics of the lips, with movement starting roughly half a second before phonation, and maximum displacement at phonation onset²¹. The fact that maxima mostly occur at vocal onset can be observed also in Fig. 2b, where the signals were aligned to their maximum value instead of the vocal onset. We take advantage of this to compare the signals of the oral block (red) with those of silent block (blue), for which we do not have the vocal onset reference. Two groups are readily identified among the silent readers (blue), one with low and the other with virtually null lips activity. Direct comparison between modalities shows that, when present, articulatory movements during silent reading are roughly two orders of magnitude smaller than those of oral reading. These results allow us to explore how articulatory processing affects silent reading, in a context of strong inhibition of vocal articulation.

Silent reading of infrequent CCVs is modulated by consonantal transitions. In this section we analyze the effects of phonation and memory on the ocular dynamics during reading. We focused on the plosive-plosive cluster, which presents two methodological advantages over the other ones (fricative-fricative



		Consonantal transition τ'	Repetition Number	Intra-word frequency
	Vocal onset delay Δ	$t(763) = 5.54, p < 10^{-4}$	$t(763) = -1.74, p = 0.082$	$F(2,763) = 1.32, p = 0.25$
	Total fixation time TFT	$t(723) = 2.13, p = 0.034$	$t(723) = -5.60, p < 10^{-4}$	$F(2,1064) = 1.82, p = 0.18$
	First pass FPRT	$t(723) = 1.78, p = 0.076$	$t(723) = -5.16, p < 10^{-4}$	$F(2,1064) = 1.25, p = 0.26$
	First fixation FFD	$t(723) = -0.094, p = 0.93$	$t(723) = 0.59, p = 0.56$	$F(2,1064) = 0.33, p = 0.57$
	Total fixation time TFT	$t(13) = 2.23, p = 0.044$	$t(1109) = -9.15, p < 10^{-4}$	$F(2,1108) = 0.10, p = 0.75$
	First pass FPRT	$t(13) = 1.60, p = 0.13$	$t(1109) = -9.01, p < 10^{-4}$	$F(2,1108) = 0.39, p = 0.53$
	First fixation FFD	$t(13) = 3.60, p = 3.22 \cdot 10^{-3}$	$t(1109) = 0.84, p = 0.40$	$F(2,1108) = 1.25, p = 0.26$

Table 1. Effects of phonology and memory on ocular variables during reading CCVs in the plosive-plosive cluster. An ANOVA test revealed no significant effect of intra-word frequency on any of the variables. **Oral reading block.** A multiple linear regression was conducted for the onset delay Δ and for each of the ocular measures, with consonantal transition and repetition number as independent variables. **Silent reading block.** To compare variables across blocks, two different tests were performed. Short term memory effects on the ocular measures was accounted for by performing a linear regression test with repetition number as the independent variable. Ocular variables (silent block) and consonantal transitions (oral block) were averaged across participants and repetitions before performing a linear regression test weighted by the statistical errors (underlined). Significance level was set to $\alpha = 0.01$ to account for multiple comparisons.

and plosive-fricative): first, acoustic features are sharply defined in the spectrogram, warranting that both Δ and τ' can be reliably identified; second, the intra-word frequency of appearance in Spanish is homogeneously low within the cluster (Fig. 1b), making the combinations comparable between themselves. The analysis for the other clusters can be found in Supplementary Figs. S1 and S2.

We used the consonantal transition τ' as the main phonatory feature of the CCVs; since participants read each CCV six times (three times per block), the repetition number was used to test for short term memory effects. Long term memory effects were examined using the frequency of appearance of the CCVs in a Spanish corpus. Statistical tests show no effects of frequency on any visual variable (last column of Table 1) as expected by the low frequency levels of the whole set of plosive CCVs (with the exception of *bta*), as sketched in Fig. 1b.

Oral reading block. We analyzed the effects of phonation and memory on the ocular variables and also on the pre-phonatory variable Δ during oral reading (Table 1, top row). For the latter, we found a significant effect of τ' on Δ , as reported in a previous work¹⁷. This means that the preparatory stages before phonation are longer for CCVs requiring longer consonantal transitions, which has been identified as a possible effect of articulatory processing¹⁷. For the ocular variables, linear regressions were conducted using consonantal transition τ' and repetition number as independent variables. A negative dependence with repetition number emerged on FPRT and TFT, and no effects were found on FFD, supporting that first fixation durations are not sensitive to habituation during oral reading. Finally, we found no effects of transitions τ' on the ocular variables. This result was expected, given the complexity of ocular dynamics during oral reading⁹, in which the eyes appear to be holding in place for many fixations “so as to not get too far ahead of the voice”²².

Oral vs. silent reading blocks. We next concentrated on the relationship between oral and silent reading, which is the main aim of this work. For this sake, we compared ocular data from the silent block with phonological data from the oral block (Table 1).

To compare cross-block variables, we first performed linear regressions for each ocular measure using repetition number as the independent variable. Consistent with the results obtained for the oral block, this revealed that FPRT and TFT systematically decrease across repetitions, while FFD presents no short term memory effects. This gives us confidence that the variables that integrate the fixations on a CCV (FPRT and TFT) are affected by habituation while the duration of the first fixation FFD is not, making the latter a good candidate for articulatory processing.

To test this, we compared the ocular variables of the silent block with the phonatory variable τ' of the oral block, averaging across participants and repetitions (Table 1, lower block of first column). This is summarized in Fig. 3, where we show each ocular variable as a function of the consonantal transitions per CCV. A strong positive correlation emerged for FFD (Fig. 3a), while for TFT and FPRT the relation did not reach significance (Fig. 3b, c).

Taken together, these analyses reveal that first fixations are not affected by habituation during the silent reading of these CCVs, and are tightly correlated with the transitions τ' . This supports that ocular dynamics are strongly modulated by phonatory features during the silent reading of infrequent combinations of letters.

Consonantal transitions measure articulatory efforts. So far, our results support that silent reading is strongly modulated by the transitions τ' between consonants. Here we connect this timing variable with the motor actions needed to produce the utterance of a CCV. Like any other speech structure, producing a CCV requires two basic motor actions: the larynx needs to be abducted to control the oscillations of the vocal folds, and the vocal tract shape is reconfigured to produce the different phonemes. We explored how the consonantal transitions τ' are related to these specific motor actions.

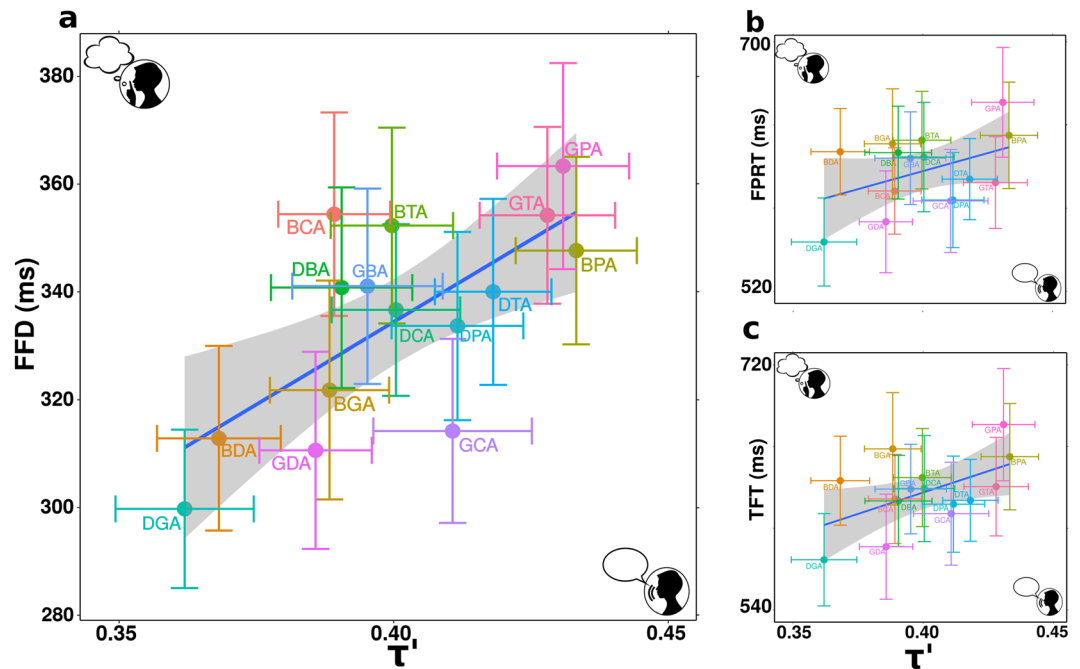


Figure 3. The silent reading of CCVs is modulated by consonantal transitions. **(a)** First fixation durations from the silent reading block as a function of transition τ' from the oral reading block. A linear regression was conducted using CCV mean values and standard errors, showing a strong positive correlation (FFD: $t(13) = 3.60$, $p = 3.22 \cdot 10^{-3}$). **(b,c)** Correlations for the other visual variables are not significant (first pass reading time FPRT: $t(13) = 1.60$, $p = 0.13$; total fixation time TFT: $t(13) = 2.30$, $p = 0.044$).

Laryngeal effort. The vocal folds are a pair of elastic membranes located at the glottis, within the larynx. The folds can be set into oscillatory motion by the passing airflow, when the lung pressure is raised over a threshold that increases with glottal abduction. Within our set of CCVs, the folds are either vibrating during the whole vocalization (*bda*, *bga*, *dba*, *dga*, *gba*, *gda*) or not vibrating during the second plosive (*bca*, *bta*, *bpa*, *dca*, *dt*, *dpa*, *gca*, *gta*, *gpa*), as evidenced by the traces of the fundamental frequency in the spectrograms of Fig. 1d. To produce the CCVs in the second group, the glottis needs to be actively abducted to prevent vibration during the middle unvoiced plosive, which involves a devoicing effort. Consonantal transitions reflect this, with a smaller mean duration for the former group, $\tau' = (38.2 \pm 0.3)10^{-2}$ than for the latter, $\tau' = (41.3 \pm 0.3)10^{-2}$ (t-test: $t(724) = -4.99$, $p < 10^{-4}$). Data is presented as mean \pm standard error (s.e.m.).

Tract effort. A CCV formed by plosives is produced by two successive occlusions that occur while the tract evolves towards the vowel *a*. This produces slight differences in the articulatory effort when consonants are interchanged. For instance, *bda* requires less effort than *dba* because in the latter, the lips' closure for the *b* occurs closer to the vowel *a*, for which the mouth is fully opened, requiring a larger vocal tract deformation. These slight vocal tract asymmetries are also reflected by consonantal transitions, with $\tau' = (37.3 \pm 0.7)10^{-2}$ for the group (*bda*, *bga*, *dga*) and $\tau' = (39.0 \pm 0.7)10^{-2}$ for the group with interchanged consonants (*dba*, *gba*, *gda*), with a trend towards significance ($t(283) = 1.79$, $p = 0.074$). Beyond this subtlety, the main variations in the tract effort arise from the anatomical differences between plosives: for instance, producing a *g* requires displacing the body of the tongue, while for a *d* only the tip is shifted. These anatomical differences can be accounted for using mathematical functions $A(x, t)$ for the cross section of the vocal tract along its length x from the entrance to the mouth.

In a previous work¹⁷ we capitalized on this mathematical description, joining together the laryngeal and vocal tract components in a single equation for the vocal effort E that reads:

$$E = E_0 + \int_0^T \int_0^L |A(x, t) - A_\Omega(x)| dx dt \quad (1)$$

The first term refers to the laryngeal component, that takes a constant value $E_0 > 0$ when a phoneme is devoiced, and $E_0 = 0$ otherwise. The second term accounts for the elastic forces produced by deformations of the vocal tract, from a neutral shape $A_\Omega(x)$ to a general shape $A(x)$, as sketched in Fig. 4a. The tract effort is obtained by integrating the section $A(x, t)$ along the duration T of a CCV, from the vocal tract entrance at $x = 0$ to the mouth exit at $x = L$. In this way, Eq. 1 provides an estimate of the total laryngeal and vocal tract articulatory actions needed to produce our speech structures. To compute the efforts of each CCV, we used vocal tract actions $A(x, t)$ and devoicing efforts E_0 from anatomical data as reported elsewhere^{17,23,24}. In Fig. 4b we show τ' as a

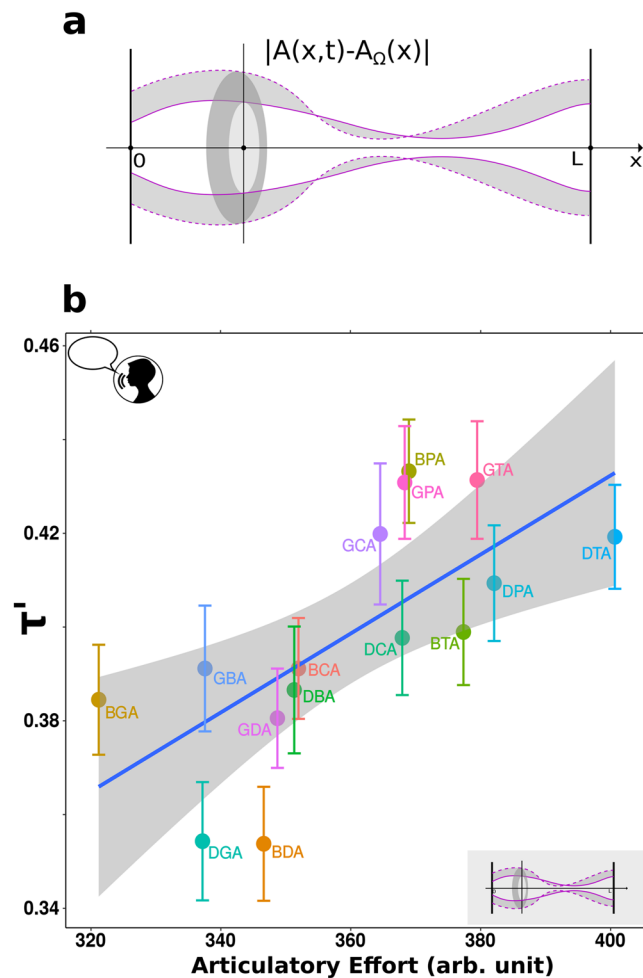


Figure 4. Computing total vocal effort from consonantal transitions. **(a)** The shape of the vocal tract is described by its cross section $A(x, t)$, where x is the distance from the vocal tract entrance ($x = 0$) to the exit of the mouth ($x = L$). The articulatory effort needed to pronounce a CCV can be estimated from the deformations of the vocal tract (shaded region) with respect to a relaxed configuration (dotted line). **(b)** The transitions τ' measured on the recorded CCVs averaged across participants, exhibit a positive relation with the modeled articulatory effort computed with Eq. 1 ($t(13) = 3.4$, $p = 4.7 \cdot 10^{-3}$).

function of the effort E for each CCV, averaged over participants and repetitions (for details see Materials and Methods, Vocal tract model). This result supports that consonantal transitions are a good estimate for the vocal effort needed to produce a CCV at laryngeal and vocal tract levels.

Taken together, the analyses of the previous sections show that decoding unfamiliar combinations of letters involves the processing of articulatory features. First fixations during silent reading are lengthened when the CCVs require a greater laryngeal and/or articulatory effort to be pronounced, even when this implies devoicing a single consonant or changing the point of occlusion in the vocal tract.

Conclusions and Discussion

The faculty of language involves the ability to switch between inner and actual speech, offering a unique opportunity to study the differences in motor processing between these operations. We capitalized on this faculty to explore the signatures of articulatory decoding during the silent reading of infrequent consonant-consonant-vowel trigrams (CCV). Our main findings are that: 1. the ocular dynamics are predicted by the consonantal transition times computed from the overt reading speech spectrogram; and 2. this transition time, in turn, correlates with the articulatory effort required to produce the utterance.

Two theories have accounted very differently on the role of the motor system during inner speech tasks. The abstraction theory supports that inner speech only activates abstract linguistic representations, independently from any articulatory mental simulation^{25,26}. On the other hand, the motor simulation theory^{27,28} describes inner speech as an activity that involves a similar motor processing than overt speech, including articulatory detail. Halfway between both theories, Oppenheim and Dell proposed a flexible abstraction hypothesis, in which inner speech operates at two levels: an abstract processing level, and one that incorporates a lower-level articulatory planning²⁹. Interestingly, our results hold for CCV structures comprising plosive-plosive transitions, but not for

fricative-fricative or plosive-fricative ones. This lack of generalization could derive from two different factors. On one side, vocalizations comprising fricatives have fuzzier spectra than the plosive-plosive ones, and therefore the associated timing variables are noisier. On the other, while the plosive-plosive combinations have a low frequency of appearance in Spanish, such frequency is higher -and more variable- for combinations including fricatives. Crucially, while no effect of frequency was found for the plosive-plosive ocular dynamics (Table 1), the same analysis reached significance for the fricatives' cluster (Supplementary Table S1). Based on these results, we hypothesize that the silent reading of frequent and infrequent combinations of letters is supported by different brain mechanisms, where a phonological decoding is necessary only for the first case. Moreover, this line of reasoning suggests that the articulatory representation mechanism could play a crucial role during the early stages of reading acquisition.

The role of speech motor representations in different cognitive processes has been widely investigated; Whitford *et al.*² showed that speech imagery is associated with an efference copy with detailed auditory properties, treating it as a kind of action. In the same direction, a groundbreaking result was reported by Anumanchipalli and colleagues³⁰, who mounted a neural decoder based on articulatory dynamics capable of synthesizing speech when a participant silently mimed sentences. These results show that investigation on inner speech has direct implications in the development of devices to restore speech³¹, and also in other fields as psychiatry, where failure in the monitoring of inner speech has been associated with verbal hallucinations in schizophrenia³². However, not so much attention has received the motor processing during silent reading. Previous research showed that silent reading activates phonological representations in the brain^{6,8}, and also that global aspects of overt speech affect the dynamics of silent reading⁷. In line with this research, our work is a quantitative in-depth study of the articulatory signatures used to decode unfamiliar letter strings during silent reading. We have shown that slight motor variations are sufficient to affect the reading dynamics and indicate that this task, as speech imagery, is assisted by vocal simulations.

Materials and Methods

Participants. Thirty native Spanish speakers (13 females and 17 males, age range 22–33, mean age 27), undergraduate and graduate students at the University of Buenos Aires with normal vision, no speech impairments and fluent reading skills completed the experiment. Participants signed a written consent form. All the experiments described in this paper were approved by the CEPI Ethics Committee of the Hospital Italiano de Buenos Aires, qualified by ICH (FDA-USA, European Community, Japan) IRb00003580, and all methods were performed in accordance with the relevant guidelines and regulations.

Stimuli and tasks. The experiment consisted of two blocks separated by a three minute break, each involving a single reading modality. Before the oral reading block, participants were instructed to say the pseudo-words as they appeared on the screen. Before the silent reading block, participants were instructed to read the sequences silently, as they normally do when reading a book. The order of the blocks was randomized.

The experimental design was identical for each block: a screen containing a sequence of 6 randomized consonant-consonant-vowel (CCV) structures was presented right after the participant fixated a red point (Fig. 1a). Once the CCVs were read, the participant pressed a key and the process restarted until completing the set (Supplementary Videos S1 to S5).

The CCVs were built using fricative and plosive consonants followed by the vowel *a* (as pronounced in *father*). We used the most common Spanish fricatives *f*, *s* and *j* (*face*, *stand*, and the Scottish *loch*) and all the Spanish voiced plosives *b*, *d*, *g* (*bay*, *dye* and *gray*) and unvoiced plosives *p*, *t*, *c* (*pay*, *tie* and *cray*). Examples of these structures are *fsa* (fricative-fricative), *sda* (fricative-plosive) and *bta* (plosive-plosive). We excluded CCVs that repeat the same consonant (i.e. *tta*). During initial trials in which these combinations were used, some participants pronounced them as a single, longer consonant. We prioritized giving our participants simple directions and therefore excluded these combinations from the set. This makes a total number of $9 \times 8 = 72$ CCVs. Here we analyzed the 15 CCVs formed by a voiced plosive (*b*, *p*, *d*) followed by any other plosive of the set (*t*, *g*, *c*, *b*, *p*, *d*). The reason is that in these combinations, vocal folds vibrate the initial voiced plosives, allowing us to determine precisely the beginning of the CCV in the spectrogram. Initial unvoiced plosives, on the contrary, leave no sound or spectrogram traces and were excluded from the analysis. Voiced plosive-plosive combinations were repeated three times per block (15 CCVs \times 3 repetitions = 45), while all other 57 CCVs were repeated randomly two or three times -adding 135 repetitions- to make a total of 180 stimuli (6 CCVs per screen \times 30 trials) per block. Data was recorded for 10800 CCV samples (30 subjects \times 2 blocks \times 180 CCVs). In order to avoid screen edge effects on the ocular variables, the first and last CCVs on the screen were discarded (Fig. 1a). The final database was composed by 7200 CCV samples.

The frequency *f* of each combination of consonants was computed from a large Spanish corpus³³ as the intra-word appearances per million words. Since the range of frequency values was large ($0 \leq f \leq 10^5$), the range was discretized in three levels: low ($f \leq 10$), medium ($10 \leq f \leq 50$) and high ($f \geq 50$). None of the CCVs are Spanish words. Courier New monospaced font was used to ensure a fixed-width of 24 pixels for every character, which corresponds to about 1 degree of visual angle. CCVs were separated by 4 blank spaces, which correspond to about 4 degrees. Stimuli were presented in black over a white background.

Eye tracking data. Visual trajectories were acquired through a desktop-mounted, video-based eye tracker (EyeLink II; SR Research Ltd., Kanata, Ontario, Canada) at a sampling frequency of 1 kHz in binocular mode (nominal average accuracy 0.5°, space resolution 0.01° RMS). The stimuli were presented on a 19-inch monitor model Samsung SyncMaster 997 MB, at an eye-monitor distance of 50 cm. Eye position was recorded at a resolution of 1024 \times 768 pixels in all tasks. The head was stabilized with a forehead rest.

Eye movements during reading include saccades, which are rapid eye movements, and fixations, where the eyes remain relatively still for about 200–300 ms. In this experiment we computed three of the most frequent variables used for measuring processing time²²: 1. First Fixation Duration (FFD), 2. First Pass Reading Time (FPRT), which is the sum of all fixation durations on the word before any other word is fixated, and 3. Total Fixation Time (TFT), the sum of all fixations durations on a word.

Acoustic and articulatory data. Sound was recorded with a commercial microphone placed at 0.3 m from the head of the participant. Lip movement was recorded using an accelerometer ADXL335 fixed to the lower lip with medical paper tape. Recordings were made in the direction of maximum amplitude of lip movement, along the perpendicular axis of the accelerometer. Sound and articulatory signals were acquired with a DAQ (USB-1608FS-Plus, Measurement Computing, Norton, Massachusetts), which ensured synchronization of both inputs and a maximum onset delay of 50 ms. Psychtoolbox library from MATLAB was used to synchronize the eye tracker and DAQ system to the computer clock.

For each CCV, we measured the following three timing variables from the audio records:

- The delay Δ between the first fixation and the vocalization onset. Vocal onsets were established at the appearance of the spectral signature of the initial consonant (noisy spot for fricatives and fundamental frequency for voiced plosives).
- The duration T of the uttered CCV.
- The transition τ . We followed the procedure described in¹⁷ to define the transitions τ in the time-frequency domain. For voiced plosive-plosive combinations, as shown in Fig. 1d, τ goes from the release of the first plosive, characterized by a voiced sound structure, to the release of the second one into the vowel a ; for plosive-fricative combinations, τ is the interval bounded by the brief broad-band plosive noise and the purely noisy fricative spot, as shown for *gfa* in Supplementary Fig. S1a; for fricative-plosive combinations, the transition is the interval between the abrupt end of the fricative and the release of the plosive into the vowel, as shown for *fca* in Supplementary Fig. S2a.

To wash out the speech rate differences between speakers, normalized transitions $\tau' = \tau/T$ were used throughout this work.

From the recorded 2700 voiced plosive-plosive combinations, we discarded those that were not fixated with the eyes, the mispronounced ones and those that contained more than one attempt to produce a CCV. We have also excluded 2 participants who mispronounced more than 45% of the CCVs throughout the experiment, leaving a dataset of 2415 CCVs. We then discarded repetition which presented fixation durations (FFD, FPRT and TFT), delay times Δ or transitions τ' out of the 95% confidence interval ($\bar{x} \pm 2s$) from the participant's mean value of each variable. A final dataset of ocular, acoustic and articulatory data of 2178 CCVs from 28 participants was used to perform the analyses throughout this work.

Vocal tract model. During a sequence of vowels and n plosive consonants, the vocal tract area $A(x, t)$ can be modelled by²³:

$$A(x, t) = \frac{\pi}{4} [\Omega(x) + q_1(t)\varphi_1(x) + q_2(t)\varphi_2(x)]^2 \prod_{k=1}^n [1 - c_k(x)m(t - t_k)] \quad (2)$$

The squared factor in Eq. 2 represents the vowel substrate, with empirical functions Ω , φ_1 and φ_2 obtained from an orthogonal decomposition calculated from MRI anatomical data. The specific features of each plosive consonant are represented by the spatial functions c_k that characterize the anatomy of the occlusion and also by the function $m(t)$ that represents the temporal dynamics of the occlusion (activation-deactivation). These functions reach the value 1 around $x = x_k$ and $t = 0$ respectively, producing a specific occlusion $A(x_k, t_k) = 0$.

This model has been tested in two ways. First, numerical integration of the vocal model produces synthetic speech samples that are indistinguishable from human speech¹⁶. Second, the model produces intelligible speech when driven by experimental vocal gestures using three detectors in the oral cavity³¹. This background helps supporting the pertinence of the model for vocal simulations.

For the CCVs used here, we set the functions $q_1(t)$ and $q_2(t)$ to evolve linearly²³ from a neutral vocal tract to the Spanish vowel a , in a time $T = 490$ s (mean CCV duration across participants). The evolution of the plosives is driven by the functions $m(t)$ that activate the vocal tract closures. To calculate the effort associated with each CCV, a range of intervals Δt between closures was explored to account for every possible degree of overlapping between both plosives¹⁷. In all cases, efforts are positively correlated with τ' in the cluster. Effort values in Fig. 4b were obtained setting an interval of $\Delta t = 200$ ms between closures (mean τ across participants) for every CCV in the cluster.

Data availability

All data generated and scripts used during this study are available in the Github repository, https://github.com/alantaitz/Data_and_Scripts_Motor_Representations.

Received: 4 September 2019; Accepted: 13 December 2019;

Published online: 02 March 2020

References

- McGuigan, A. F. J., Camacho, E. O., Hardyck, C. D., Petrinovich, L. F. & Delbert, W. Feedback of Speech Muscle Activity during Silent Reading: Rapid Extinction. *Science* (80-). **157**, 579–581 (1967).
- Whitford, T. J. *et al.* Neurophysiological evidence of efference copies to inner speech. *Elife* 1–23, <https://doi.org/10.7554/eLife.28197.001> (2017).
- Cheung, C., Hamiton, L. S., Johnson, K. & Chang, E. F. The auditory representation of speech sounds in human motor cortex. *Elife* **5**, 1–19 (2016).
- Martin, S. *et al.* Decoding inner speech using electrocorticography: Progress and challenges toward a speech prosthesis. *Front. Neurosci.* **12**, 1–10 (2018).
- Laubrock, J. & Kliegl, R. The eye-voice span during reading aloud. *Front. Psychol.* **6**, 1–19 (2013).
- Abramson, M. & Goldinger, S. D. What the reader's eye tells the mind's ear: Silent reading activates inner speech. *Percept. Psychophys.* **59**, 1059–1068 (1997).
- Filik, R. & Barber, E. Inner speech during silent reading reflects the reader's regional accent. *Plos One* **6** (2011).
- Ashby, J., Sanders, L. D. & Kingston, J. Skilled readers begin processing sub-phonemic features by 80 ms during visual word recognition: Evidence from ERPs. *Biol. Psychol.* **80**, 84–94 (2009).
- Ashby, J., Yang, J., Evans, K. H. C. & Rayner, K. Eye movements and the perceptual span in silent and oral reading. *Attention, Perception, Psychophys.* **74**, 634–640 (2012).
- Dehaene, S. *Reading in the brain: The new science of how we read.* (Penguin, 2009).
- Fernández, G., Shalom, D. E., Kliegl, R. & Sigman, M. Eye movements during reading proverbs and regular sentences: The incoming word predictability effect. *Lang. Cogn. Neurosci.* **29**, 260–273 (2014).
- Kliegl, R., Grabner, E., Rolfs, M. & Engbert, R. Length, frequency, and predictability effects of words on eye movements in reading. *Eur. J. Cogn. Psychol.* **16**, 262–284 (2004).
- Story, B. H., Titze, I. R. & Hoffman, E. A. Vocal tract area functions from magnetic resonance imaging. *J. Acoust. Soc. Am.* **100**, 537–54 (1996).
- Story, B. H. & Bunton, K. Relation of vocal tract shape, formant transitions, and stop consonant identification. *J. Speech. Lang. Hear. Res.* **53**, 1514–28 (2010).
- Assaneo, M. F., Nichols, J. I. & Trevisan, M. A. The Anatomy of Onomatopoeia. *Plos One* **6**, e28317 (2011).
- Assaneo, M. F. *et al.* Exploring the anatomical encoding of voice with a mathematical model of the vocal system. *Neuroimage* **141**, 31–39 (2016).
- Taitz, A., Shalom, D. E. & Trevisan, M. A. Vocal effort modulates the motor planning of short speech structures. *Phys. Rev. E.* **052406**, 1–7 (2018).
- Clifton, C. Jr., Staub, A. & Rayner, K. Eye movements in reading words and sentences. in *Eye Movements* 341–371 (Elsevier, 2007).
- Vasishth, S., von der Malsburg, T. & Engelmann, F. What eye movements can tell us about sentence comprehension. *Wiley Interdiscip. Rev. Cogn. Sci.* **4**, 125–134 (2013).
- Goldstein, L., Byrd, D. & Saltzman, E. The role of vocal tract gestural action units in understanding the evolution of phonology. in *Action to Language via the Mirror Neuron System* 215–249 (Cambridge University Press, 2006).
- Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–32 (2013).
- Rayner, K. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychol. Bull.* **124**, 372–422 (1998).
- Story, B. H. A parametric model of the vocal tract area function for vowel and consonant simulation. *J. Acoust. Soc. Am.* **117**, 3231 (2005).
- Story, B. H. Phrase-level speech simulation with an airway modulation model of speech production. *Comput. Speech Lang.* **27**, 989–1010 (2013).
- MacKay, D. G. Constraints on theories of inner speech. in *Auditory imagery* 133–162 (Psychology Press, 2014).
- Dell, G. S. & Repka, R. J. Errors in Inner Speech. In *Experimental Slips and Human Error* 237–262 (Springer, 1992). https://doi.org/10.1007/978-1-4899-1164-3_10
- Levelt, W. J. M. *Speaking: From intention to articulation.* **1**, (MIT press, 1993).
- Postma, A. & Noordanus, C. Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Lang. Speech* **39**, 375–392 (1996).
- Oppenheim, G. M. & Dell, G. S. Motor movement matters: the flexible abstractness of inner speech. *Mem Cogn.* **38**, 1147–1160 (2010).
- Anumanchipalli, G. K., Chartier, J. & Chang, E. F. Speech synthesis from neural decoding of spoken sentences. *Nature* **568**, 493–498 (2019).
- Assaneo, M. F., Butavand, D. R., Trevisan, M. A. & Mindlin, G. B. Discrete anatomical coordinates for speech production and synthesis. *Front. Commun.* **4**, 1–13 (2019).
- Brébion, G. *et al.* Impaired self-monitoring of inner speech in schizophrenia patients with verbal hallucinations and in non-clinical individuals prone to hallucinations. *Front. Psychol.* **7**, 1–12 (2016).
- Cuetos, F., Glez-Nosti, M., Barbón, A. & Brysbaert, M. SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica* **32**, 133–143 (2011).

Acknowledgements

We thank Juan Kamienkowski for his time and willingness during the experiments. The research reported in this work was partially funded by the National Scientific and Technical Research Council (CONICET) and the University of Buenos Aires (UBA).

Author contributions

A.T., M.F.A., D.E.S. and M.A.T. designed the research and wrote the manuscript. A.T. acquired the data. A.T., D.E.S. and M.A.T. analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-59199-6>.

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

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020