# MONITORING CV SYLLABLES VERSUS PHONEMES IN INTERNAL SPEECH

Pierre A. Hallé,[a] Juan Segui,[b] and Laura Manoiloff[c]

[a] LPP (CNRS-Paris 3), [b] MC2-lab (INSERM-Paris 5), [c] Univ. Nacional de Córdoba (Argentina)
pierre.halle@sorbonne-nouvelle.fr, juan.segui@parisdescartes.fr, lmvmanoiloff@gmail.com

## ABSTRACT

English and French listeners have been found to monitor faster utterance-initial CV syllables than C consonants in overt speech. We report a similar "syllable advantage" with Spanish-speaking Argentinean participants monitoring for inner speech. They were faster to detect CV syllables than C consonants at the beginning of the names of the pictures they were presented with (Experiment 1). This syllable advantage in internal monitoring resisted to adding "foil pictures," whose name's first syllable differed from the target syllable by the vowel only (Experiment 2), a manipulation logically more detrimental to syllable than phoneme detection. The foils induced a more cautious strategy across the board (longer RTs) but the syllable advantage was maintained. Our results converge with prior studies in revealing intriguingly parallel perceptual effects in overt and inner speech, suggesting that both are processed using a common abstract code, possibly based on syllabic gestural scores.

**Keywords**: Inner-speech, internal-monitoring, syllable-detection, phoneme-detection.

## 1. INTRODUCTION

Planning spoken utterances requires a series of processing operations on representations retrieved at the conceptual, syntactic, lexical, morphemic, and phonological levels, performed quickly and out of conscious control. Among these operations, the phonological processing stage (the "phonological encoding" stage in Levelt's model of speech production: [11, 12]) is essential to achieve the ultimate goal of the system: speak out an intelligible message.

Phonological encoding takes as input the information of the lexical representation currently selected (the "lexeme"), which specifies a word-form, presumably on different dimensions of information. For example, [12] proposes that segmental and metrical information are initially specified separately in lexical entries and then recombined by a "phonological encoder" module into syllabic slots, which the encoder fills up with segmental content, after applying context-dependent phonological processes where needed. The resulting code is further converted into syllable-sized "articulatory scores," which are the input to the "phonetic plan" that generates the "motor plan" to be executed, resulting in a complex time-structured set of articulatory gestures ("articulatory map") producing the speech output. We refer here to Levelt's model for the sake of clarity but other models also assume processing stages whereby an articulatory map is elaborated from lexical and contextual phonological information ([4, 9, 10]).

Speech production models assume the existence of devices that can *monitor* internal speech during speech production for appropriateness and errors. Two main related issues may be raised concerning internal monitoring: its functional locus and the nature of the representations examined during internal monitoring. Some researchers propose internal monitoring is achieved within the phonological encoding system [20]. Others assume it is achieved via a single speech comprehension system processing either external or internal speech. The labels "production monitor" vs. "perception monitor" have been proposed by [20] to account for these two viewpoints. As for the code examined for internal monitoring, it is probably more abstract than that used for the motor plan produced by the phonetic plan. However, its precise nature requires further specification.

In a seminal work, Wheeldon and Levelt [24] used internal phoneme and syllable monitoring tasks to study the properties of internal speech. In their experiments, Dutch-English bilingual subjects had to silently generate the Dutch translation of an auditorily presented English word and to monitor the internal speech they generated –the Dutch translation– for a pre-specified phoneme or syllable target. Their results supported the *perceptual-loop* hypothesis, which proposes that the input to the *phonetic plan* is incrementally fed to *the* speech comprehension system: this single system would monitor internal, planned speech just like it monitors external, overt speech, operating on the same information flow. The results in [24] also suggest that the code circulating in the perceptual loop be more abstract than the code used in the phonetic plan, because its monitoring is insensitive to concurrently planned articulation in an *articulatory suppression* condition.

A prediction derived from this hypothesis is that presumably perception-specific effects found with overt speech might be found in internal monitoring of inner speech. In agreement with this prediction, Özdemir and Levelt [17] observed, in an internal

phoneme monitoring experiment on picture names in Dutch, that monitoring latencies depend on the position of the target phoneme relative to the uniqueness point (UP) of the word generated internally. For example, /l/ was detected faster in *zadel* 'saddle' (UP=/d/) than *vogel* 'bird' (UP=/e/) than *ketel* 'kettle' (UP=/l/). This "uniqueness point effect" in internal monitoring parallels the results typically obtained in overt speech perception ([15, 8, 18]) and is thus in line with the prediction made by the perceptual-loop hypothesis.

In this paper, we focus on another effect found in overt speech perception and examine whether it can also be observed in internal speech monitoring. This test can be viewed as a diagnostic tool for the perceptual-loop hypothesis. The effect at stake has been found in several studies in French and English: syllables are detected more quickly than phonemes (consonant or vowel) in word-initial position. It was initially found in English for CVC syllables ([19, 7]). [16] later showed that the effect with CVC syllables depends on the design of the experimental lists, in particular on whether lists contain catch trials with foil syllables and/or foil phonemes. Segui et al. ([21]) used no foils and found a robust advantage for CV syllables over C consonants in French: /ba/ was detected faster than /b/ in *bateau* /bato/ 'boat' as well as in /ba/-initial *nonwords*. [21] interpreted the advantage of the syllable over the phoneme as due to the immediate availability in prelexical perception, at least for stop-vowel CV sequences, of both C and V, that is, of the CV sequence as a whole, based on the groundbreaking (at the time) findings in [1, 22].

The issue we address here is of whether the faster monitoring of CV syllables than C phonemes is also found when monitoring internal (inner) speech.

We used pictures whose name in Spanish (same rhythmic class as French) was two- or three-syllable long and began with a CV syllable. Spanish-speaking Argentinean participants had to monitor for the word-initial C or CV of these pictures' names. The design of Experiment 1 closely followed [21], except that participants had to monitor for syllables or phonemes in internally generated rather than externally presented word-forms.

## 2. EXPERIMENT 1

### 2.1. Methods

***Participants.*** Eighty-one students at National University of Cordoba, Psychology Department, aged 18–26 years, participated voluntarily in the experiment. All were Argentinean native speakers of Spanish, with normal or corrected-to-normal vision and no known language, speech, or hearing disorder.

***Materials.*** Twenty black-on-white drawings of simple common objects, selected from the set described in [3], served as test picture stimuli. Their name began with CV syllables. There were 10 word-initial target phonemes (/b, d, g, p, t, k, v, f, m, n /) and 13 word-initial target syllables (/bo, de, ga, pa, pe, pi, ti, ko, va, ve, fo, mo, ni /). Naming agreement for the selected experimental pictures was above 80% according to the norms established for Argentinean Spanish ([13]). There is no Argentinean Spanish lexical database. We thus collected subjective frequency 1–5 ratings (5 for frequent) from 25 native Argentinean speakers who did not participate in the experiments. The average subjective frequency of the 20 test picture names was 1.7 (range 1.1–2.5).

***Design.*** The set of 20 test pictures was divided into two subsets, as balanced as possible in terms of picture name subjective frequency, number of syllables and phonemes, and broad type of onset phoneme. Each subset comprised 10 test pictures plus 82 (subset 1) or 79 (subset 2) filler pictures. The pictures of each subset were blocked by either phoneme or syllable target. Each block contained about 8 times more filler than test items. Each subset contained five phoneme-target blocks, or six or seven syllable-target blocks. The 81 participants were randomly assigned to one of two groups. Both groups received first subset 1 and then subset 2. In one group (n=40), participants had to detect phonemes in subset 1 and syllables in subset 2. In the other group (n=41), participants had to detect syllables in subset 1 and phonemes in subset 2. The experimental design was thus counterbalanced across subjects for target type order, though not for subset presentation order. That is, each participant saw each of the 181 pictures only once, for either phoneme or syllable monitoring. Within each block, pictures were presented in a random order ensuring that test items did not appear in block-initial or block-final position, were separated by at least three filler items, and were not semantically or phonologically related with their flanking filler items. Within each subset, blocks were presented in a different random order for each subject.

***Procedure.*** Each phoneme- or syllable-target block was introduced with the oral specification (in Spanish) of the target to monitor for, illustrated by three words. For example, the phoneme-target /p/ was specified as "[pə] *como en 'pera', 'papa' o 'pincel'*." This was followed by a prompt asking the participant to launch the sequence of trials of the current block in pressing the spacebar. Each trial comprised the following sequence of visual stimuli

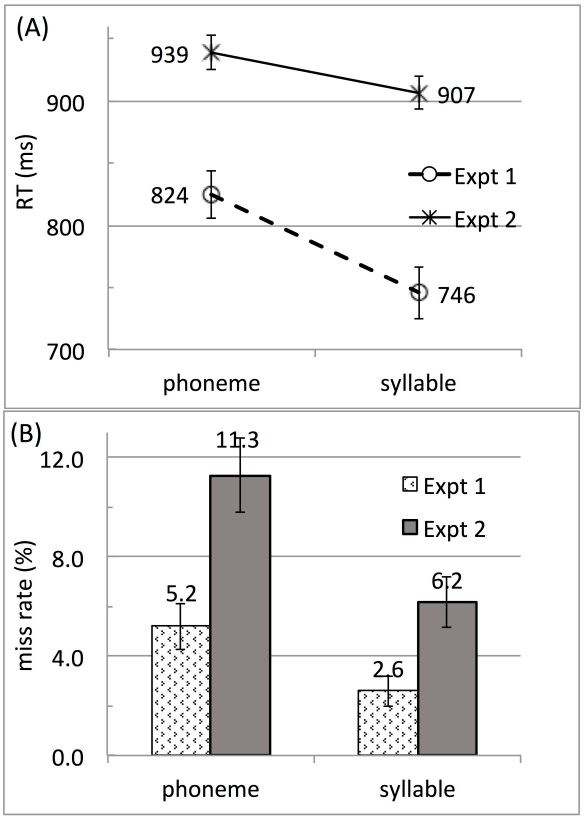
displayed at the center of the screen: "##" displayed for 2 s as a visual fixation stimulus followed by the picture in black on white within a 7 × 7 cm white square, displayed until the participant responded or for a maximum of 2 s (i.e., responses slower than 2 s were timed out). The inter-trial interval was one second. One experimental session typically lasted about 30 minutes. Participants were comfortably seated in a dimly lit quiet room. The experiment was run on a personal computer, using the DMDX experimentation software ([6]). Participants were instructed to orient their gaze at the center of the screen where the stimuli would be displayed. They were told they would be presented with series of pictures, each series associated with a phoneme- or a syllable-target to detect in the picture's name in word-initial position, and had to respond with a button press, as quickly and accurately as possible, if and only if the name of the object began with the target specified for the current series. Response times (RTs) were measured from the onset of picture presentation.

### 2.2. Results and discussion

We first computed the by-item and by-subject miss rates. We excluded eight participants with more than 20% miss rate (three in the phoneme-first group, five in the syllable-first group). No items were discarded using the 20% criterion. This left 38 and 35 participants in the phoneme- and syllable-first group, respectively. For RTs, no further filtering was applied than the 2 s time-out (no anticipated responses). RT and miss-rate data are shown in Fig. 1AB. As can be seen, syllables were detected faster and missed less often than phonemes. This was confirmed by linear mixed model analyses (with R) on the RT and miss rate data. The models best fitting the data included as fixed effects Group (phoneme- vs. syllable-first), Target (phoneme vs. syllable), and their interaction, and, as random effects, intercept and slope on Target for subjects, and intercept for items. For RTs, Target was significant, $p<.00001$, and neither Group nor Group × Target were significant ($p=.59$ and $p=.91$, respectively). For miss rates, Target was significant, $p=.021$, and neither Group nor Group × Target were significant ($p=.51$ and $p=.31$, respectively).

A possible concern with these results is that phoneme detection might have been challenged compared to syllable detection. First, within phoneme-target blocks, the word-initial consonant of 1.4 filler items (out of eight in average) differed from that of the target by a single distinctive feature, generally the place feature. Such filler items may be viewed as foils. The syllable-target blocks did not contain comparable foils: for a given CV target, the corresponding syllable-target block contained no CV' filler item. Second, two phoneme-target blocks contained C-onset test items differing by the following vowel: the /p/-target block contained /p/-items with three different vowels (/a, i, e/); likewise the /v/-target blocked contained *vela* and *vaca*. As shown by [23], consonant detection is slowed down by the "uncertainty as to the identity of the vowel following the target consonant." This might apply to inner speech. Although such uncertainty was limited to two blocks out of ten, it might have been sufficient to bias participants toward a general slowing down of phoneme detection latencies. To sum up, the materials of Experiment 1 might have disfavored, however slightly, phoneme detection compared to syllable detection, assuming, again, that similar contextual and foil effects occur in internal and external monitoring. To test for this possibility we ran a control experiment in which CV detection would presumably be substantially disfavored by the presence of CV' foils in CV-target blocks.

**Figure 1**: Experiments 1-2: (a) RT, and (b) miss rate data.




## 3. EXPERIMENT 2: CV' FOILS

Thirty CV' foils were added to the 20 CV target-bearing test items within CV-target blocks. Foil and test items, in the 3:2 ratio, shared their word-initial consonant and differed only from the subsequent vowel on. As suggested by [16], this should substantially slow down CV detection.

### 3.1. Methods

***Participants.*** Fifty-two students at National University of Cordoba, Psychology Department, aged 18-26 years, all Argentinean native speakers of Spanish, participated in Experiment 2. None had participated in Experiment 1 or reported a language disorder.

***Materials and design.*** The same 20 test pictures as in Experiment 1 were dispatched into the same two subsets and associated with the same 10 phoneme or 13 syllable targets as in Experiment 1, but 30 "foil pictures" were added. Their name's first syllable differed from the target syllable by the vowel only (e.g., picture of a *banana* in the /bo/-target block); 15 such CV' foils were added in each subset. No foils were added to the phoneme-target blocks. The experimental design was counterbalanced across subjects for target type order: one group detected first phonemes then syllables and the other group first syllables then phonemes in subsets 1 and 2.

***Procedure.*** It was the same as in Experiment 1.

### 3.2. Results

We first computed the by-item and by-subject miss rates. We excluded 5 participants with more than 20% miss rate (3 in the phoneme-first group, 2 in the syllable-first group). No items were discarded on the 20% criterion. This left 23 and 24 participants in the phoneme- and syllable-first group, respectively. All the available RT data were retained, as in Experiment 1. Syllables were still detected faster and missed less often than phonemes, although the effect was weaker for RTs (Figure 1AB), as shown by linear mixed model analyses.

The models best fitting the data included as fixed effects Group, Target, and their interaction, and, as random effects, intercept for subjects and items. For RTs, Target was significant, $p$=.035; neither Group nor Group × Target were significant ($p$=.28 and $p$=.81, respectively). For miss rates, Target was significant, $p$=.0027. Group was also significant, $p$=.0071, but Group × Target was not, $p$=.95.

The differences between the two experiments, as shown in Fig. 1, called for a further statistical analysis bearing on both, with the fixed effect Experiment (1 vs. 2) in addition to Target and Order. The random effects were intercept for subjects and items. For both RTs and miss rates, Target was significant, indicating faster syllable- than phoneme-detection across Experiments. Experiment was significant too, $p$s<.00001, reflecting shorter RTs and lower miss rates overall in Experiment 1 than 2 (785 < 923 ms; 3.9 < 8.8% misses). The Target × Experiment interaction was significant for RTs, $p$=.014, but not for miss rates, $p$=.19: the Target effect was stronger in Experiment 1 than Experiment 2 for RTs (78 > 32 ms) but not for miss rates.

To sum up, although introducing foil pictures with CV' names resulted in a general slow down of detection latencies for both syllables and phonemes, but the "syllable advantage" was still observed.

## 4. GENERAL DISCUSSION

Across two experiments, we found a robust advantage for CV syllables over C phonemes in terms of internal monitoring latencies, even in a CV' foil condition most detrimental for CV detection. These results mirror those obtained in [21] for external monitoring. Our data thus support the perceptual loop theory, whereby a single "comprehension system" processes either overt or inner speech ([11, 12, 24, 17]), adding to the prior findings of uniqueness point [17] or syllabic [24] effects common to overt and inner speech. Moreover, we found a strong foil effect: foils induce longer RTs. This effect, not yet documented in internal monitoring (as far as we know) is a typical one in external monitoring ([16]).

The perceptual loop theory implies that the same code be used to parse and monitor, at a phonological level, both overt and inner speech. We propose, following [12], that the code for inner speech –the output of the phonological encoder– consists in syllabic gestural scores. Although it incrementally feeds the phonetic plan, it is more abstract than the code used in the end-product motor plan, as suggested by [24] and [14]. In overt speech perception, the auditory input therefore must also be encoded into syllabic gestural scores (contra, e.g., [5]). Note that, in the present study, we limited ourselves to CV syllables: more complex syllables are left for future research.

We finally turn back to the syllable advantage effect. The original account offered by [21] was that the first available percept in a CV-initial utterance is CV, not C, at least for stop consonants. It was based on previous work by Stevens and Blumstein ([22, 1]) showing that a very short fragment of signal extracted from stop release (~40 ms) is sufficient to hear both C and V *at the same time*. [21] proposed that CV is first identified as a whole, whereas C can only be identified from the reanalysis of CV. This account relied heavily on the phonetic perception of *stop* consonants. On our novel proposition of a common syllabic gestural though abstract code (see [2]), the original account by [21] obviously needs be updated, yet not deeply revised. The notion that the processing system first accesses CVs as whole units, then must unpack them to identify C consonants (yet not restricted to stops) remains valid.

## 5. REFERENCES


[1] Blumstein, S., Stevens, K. 1980. Perceptual invariance and onset spectra for stop consonants in different vowel environments. *Journal of the Acoustical Society of America* 67, 648–662.

[2] Browman, C., Goldstein, L. 1992. Articulatory Phonology : An overview. *Phonetica* 49, 155–180.

[3] Cycowicz, Y., Friedman, D., Rothstein, M., Snodgrass, J. 1997. Picture naming by young children: Norms for name agreement, familiarity, and visual complexity. *Journal of Experimental Child Psychology* 65, 171–237.

[4] Dell, G. 1986. A spreading activation theory of retrieval in sentence production. *Psychological Review* 93, 283–321.

[5] Dumay, N., Content, A. 2012. Searching for syllabic coding units in speech perception. *Journal of memory and language* 66, 680–694.

[6] Forster, K., Forster, J. 2003. DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers* 35, 116–124.

[7] Foss, D., Swinney, D. 1973. On the psychological reality of the phoneme: perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior* 12, 246–257.

[8] Frauenfelder, U., Segui, J., Dijkstra, T. 1990. Lexical effects in phonemic processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance* 16, 77–91.

[9] Garrett, M. 1975. The analysis of sentence production. In: Bower, G. (ed.), *The psychology of learning and motivation*. New York: Academic Press, 133–177.

[10] Hickok, G. 2014. Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience* 29, 52-59.

[11] Levelt, W. 1989. *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

[12] Levelt, W., Roelofs, A., Meyer, A. 1999. A theory of lexical access in speech production*. Behavioral and Brain Sciences* 22*, 1-75.*

[13] Manoiloff, L., Arstein, M., Canavoso, M., Fernandez, L., Segui, J. 2010. Expanded Norms for 400 Experimental Pictures in an Argentinian Spanish-Speaking Population. *Behavior Research Methods, Instruments and Computers* 42, 452–460.

[14] Manoiloff, L., Segui, J., Hallé, P. 2015. Subliminal repetition primes help detecting phonemes in a picture: Evidence for a phonological level of the priming effects. *Quarterly Journal of Experimental Psychology* 69, 24–36.

[15] Marslen-Wilson, W. 1990. Activation, competition, and frequency in lexical access. In: Altman, G. (ed.), *Cognitive models of speech processing*. Cambridge: MIT Press, 148–172.

[16] Norris, D., Cutler, A. 1988. The relative accessibility of phonemes and syllables. *Perception and Psychophysics* 43, 541–550.

[17] Özdemir, R., Roelofs, A., Levelt, W. 2007. Perceptual uniqueness point effects in monitoring internal speech. *Cognition* 105, 457-465.

[18] Pitt, M., Samuel, A. 1995. Lexical and sublexical feedback in auditory word recognition. *Cognitive psychology* 29, 149–188.

[19] Savin, H. 1970. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior* 9, 295–302.

[20] Schiller, N., Jansma, B., Peters, J., Levelt, W. 2006. Monitoring metrical stress in polysyllabic words. *Language and Cognitive Processes* 21, 112–140.

[21] Segui, J., Frauenfelder, U., Mehler, J. 1981. Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology* 72, 471–477.

[22] Stevens, K., Blumstein, S. 1978. Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America* 64, 1358–1368.

[23] Swinney, D., Prather, P. 1980. Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts. *Perception and Psychophysics* 27, 104–110.

[24] Wheeldon, L., Levelt, W. 1995. Monitoring the time-course of phonological encoding. *Journal of Memory and Language* 34, 311–334.