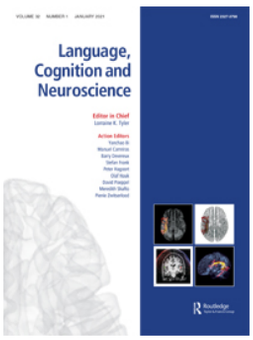

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/plcp21

Monitoring internal speech: an advantage for syllables over phonemes?

Pierre Hallé, Laura Manoiloff, Jiayin Gao & Juan Segui

To cite this article: Pierre Hallé, Laura Manoiloff, Jiayin Gao & Juan Segui (2022) Monitoring internal speech: an advantage for syllables over phonemes?, Language, Cognition and Neuroscience, 37:1, 22-41, DOI: 10.1080/23273798.2021.1941146

To link to this article: https://doi.org/10.1080/23273798.2021.1941146

LANGUAGE, COGNITION AND NEUROSCIENCE
2022, VOL. 37, NO. 1, 22–41
https://doi.org/10.1080/23273798.2021.1941146

Routledge
Taylor & Francis Group

REGULAR ARTICLE

Monitoring internal speech: an advantage for syllables over phonemes?

Pierre Hallé [a,b], Laura Manoiloff[c], Jiayin Gao[d] and Juan Segui[b]

[a]Laboratoire de Phonétique et Phonologie (CNRS-Paris 3 and EFL Labex); [b]Laboratoire Mémoire et Cognition (CNRS-Paris 5 and EFL Labex); [c]Cognitive Psychology of Language and Psycholinguistics Research Group, Center of research of the Faculty of Psychology (CIPSI), National University of Córdoba, Córdoba, Argentina; [d]Department of Linguistics and English Language, University of Edinburgh, Edinburgh, UK

ABSTRACT

Listeners generally detect syllables faster than phonemes in overt speech. This "syllable advantage" holds robustly for utterance-initial CV vs. C targets [Segui et al., 1981. Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, *72*(4), 471–477]. We report a syllable advantage when monitoring *inner* speech. Spanish-speaking Argentinian participants presented with pictures were faster and more accurate at detecting CV than C targets at the beginning of the pictures' names. This CV over C advantage maintained, although substantially weakened, after adding CV' foils in CV-target trials, a manipulation logically more detrimental to CV- than C-detection. Our results converge with previous studies showing intriguing parallelisms between overt and inner speech perception and processing, supporting a *restricted* version of Levelt's perceptual-loop hypothesis. We discuss what common basic units of processing could be, borrowing from the articulatory phonology framework and its proposal of a "common currency" between speakers and listeners.

ARTICLE HISTORY
Received 28 February 2020
Accepted 4 June 2021

KEYWORDS
Inner speech; internal monitoring; syllable detection; phoneme detection; articulatory phonology

1. Introduction

In the last decades, "inner speech," in its various forms, has been the focus of many behavioural, neuro-imagery, electrophysiological, and neuropsychological studies, as well as many speculations. Inner speech varies along several dimensions (see Grandchamp et al., 2019), among which the *condensation* dimension is of interest in the present study: from "condensed" to "expanded" (Alderson-Day & Fernyhough, 2015; Fernyhough, 2004; Grandchamp et al., 2019; also see Langland-Hassan & Vicente, 2018). That is, internally generated speech may vary from highly abstract (containing minimal semantic and structural linguistic information) to quite concrete, down to phonetic detail, or even down to indexical detail. Inner speech has also been viewed, somewhat metaphorically, as "a truncation of overt speech" (Perrone-Bertolotti et al., 2014, p. 227), that is, overt speech interrupted at some production stage from conceptualisation to articulation. On this view, interruption-level would follow degree of condensation. Our study bears on the presumably fully expanded inner speech built-up in the latest production-stages.

1.1. Commonalities between overt and inner speech

Are there similarities in the way we perceive, process, or monitor uninterrupted overt speech and late production stage inner speech? This question initially motivates the present study. A "yes" answer is suggested by several experimental results. For example, the verbal transformation effect, whereby a rapidly repeated speech utterance may undergo perceptual change (Warren, 1961) (e.g. /psə psə psə … / > /səp səp səp … /) is observed with both overt and inner speech (Reisberg et al., 1989; Sato et al., 2004; Smith et al., 1995). Tongue twister recitations show similar lexical bias effects on overt and inner slips as well as largely similar phonemic similarity effects (Corley et al., 2011; Oppenheim & Dell, 2010; but see Oppenheim & Dell, 2008). Overt and inner speech also seem to engage similar brain activity, though with notable either qualitative or quantitative differences. For comprehensive reviews, see Perrone-Bertolotti et al. (2014, pp. 223–227) and Price (2012, pp. 830–832). Some of the differences may reflect inhibition of articulation execution during inner speech (Basho et al., 2007), some others reflect auditory perception of one's own speech during overt speech. More

CONTACT Pierre Hallé pierre.halle@sorbonne-nouvelle.fr Laboratoire de Phonétique et Phonologie, ILPGA, 19 rue des Bernardins, Paris 75005, France
Supplemental data (Appendices A-C) for this article can be accessed at https://doi.org/10.1080/23273798.2021.1941146

generally, stronger activations are found in motor-related brain areas (IFG, insula, etc.) for overt than covert speech. Similarly, EMG potentials (e.g. for orofacial muscles) are much weaker during inner than overt speech (Faaborg-Andersen, Edfeldt, & Nykøbing, 1958; Sokolov, 1972; Vanderwolf, 1998, p. 128). They might however be available to sensory-motor perception.

These results, among others, suggest there are indeed commonalities in the perception and processing of overt and inner speech, at least for the kind of inner speech that is fully expanded. A very parsimonious formalised account of such commonalities is proposed by Levelt in his "Blueprint for the Speaker" or "Lemma model" speech production model (Levelt, 1989): the "perceptual loop hypothesis." Levelt proposed that both internal and overt speech are processed and monitored by a single common "speech comprehension system" (SCS), with "internal" (or "inner") speech referring to some "almost ready to go" internal representation of a planned utterance before it is spoken, and overt speech referring to some already spoken utterance. Inner speech would feed the SCS via an "inner loop" (hence the "inner" label), whereas external speech (possibly one's own speech) would feed the SCS via an "outer loop." This immediately raises a representational issue: If overt speech and inner speech undergo similar speech processing, they must share *common representations*. Shedding light on what could be the form of such representations is an important goal of the present study. With this goal in mind, our study takes as theoretical framework the rather well defined model of speech production planning proposed decades ago by Levelt (Levelt, 1989; Levelt et al., 1999; for updates and reinterpretations, see the *Thirty years of "Speaking"* special issue of *Language, Cognition and Neuroscience*: Meyer et al., 2019). Since, over the years, the Lemma model evolved, in particular with respect to the kind of inner speech inputted to the inner loop, we first sketch out the Lemma model and its evolution before turning to the rationale of the present study. In the course of the Lemma model's overview, we also mention current criticisms bearing on, if not the Lemma model's main features, at least its specific "perceptual loop hypothesis" component (e.g. Gauvin & Hartsuiker, 2020).

1.2. The Lemma model

Levelt's Lemma model (Levelt, 1989: "Speaking"), as all the models of speech production planning (e.g. Dell, 1986; Hickok, 2014a, 2014b), describes a series of operations retrieving the representations required to design a spoken utterance at the conceptual, syntactic, morpholexical, phonological, and phonetic levels. These operations are performed quickly and out of conscious control. The end product is, ideally, the audible articulation of the message intended by the speaker, that is, the phonetic form of the message. Among the operations mentioned above, the phonological processing stage is essential to achieve the ultimate goal of the system: speak out an intelligible message. Levelt (1989) called this stage "phonological encoding" in his original "Blueprint" model, and later distinguished in it a final phonetic encoding subpart (Levelt et al., 1999; see Laganaro, 2019, for more detail). In the following, we focus on phonological and phonetic encoding.

1.2.1. Phonological and phonetic encoding

It is generally assumed that phonological encoding takes as input the information of the currently selected lexical representations (the "lexemes" in Levelt's formulation). The word form specified in a lexical representation presumably combines different dimensions of form information in a complex phonological representation. Most theories of phonological encoding distinguish segmental and metrical information (Dell, 1988; Levelt, 1989; Meringer & Mayer, 1895; Shattuck-Hufnagel, 1979). The motivation for postulating separate segmental and metrical specifications is that syllabification takes place at the prosodic word level (Nespor & Vogel, 1986), not at the level of the lexical word in isolation. For example, although *demand* could be syllabified as *de.ˈmand* in isolation, *demand it* (a clitic group or prosodic word) must be syllabified as *de.ˈman.dit* (Levelt & Wheeldon, 1994). Levelt and Wheeldon (1994) explain the way this segmental vs. metrical distinction could be implemented. They propose that segmental and metrical information are initially specified separately in lexical entries, then recombined by the "phonological encoder" module – after the application of context-dependent phonological processes, where needed – into syllabic slots (metrical information), which the encoder fills up with segmental content and later converts into syllable-sized "articulatory scores." The resulting articulatory phonology representation – a sequence of syllabic gestural scores – constitutes, in Levelt's (1989) formulation, the "phonetic plan," which he also called "inner speech." The motivation for this latter, convenient label is that the phonetic plan serves as input to the "inner loop," which feeds the SCS, common to inner and outer speech, in which processing and monitoring would take place (the "perceptual loop hypothesis"). The phonetic plan must also be the input to a further motor programme elaboration stage, which is left undetailed in Levelt's model but is presumably rather complex (see Laganaro, 2019). Indeed, it

cannot be a simple translation and concatenation of syllabic gestural score specifications and attached task dynamics (Kelso et al., 1986; Saltzman & Munhall, 1989), since context-dependent task dynamics for sentence-size utterances need be computed. The end product of the motor programme elaboration stage should be, logically, the representation closest to phonetic specification. The motor programme is eventually executed to generate the complex time-structured set of articulatory gestures (or "articulatory map") producing the speech output. Other models than the Lemma model also assume some processing stage whereby an articulatory map is elaborated from lexical and contextual phonological information (Dell, 1986, 1988; Garrett, 1975; Hickok, 2014a, 2014b; Shattuck-Hufnagel, 1979).

1.2.2. Monitoring one's upcoming speech

Two main related issues can be raised concerning internal monitoring: first, its functional locus and second, the nature of the representation(s) examined during internal monitoring. At this point, we must consider the possibility that "monitoring" may take various forms. In particular, monitoring for errors may be functionally different from monitoring for some specified linguistic property. We return to this important qualitative distinction in the following. Concerning the first "locus" issue, some researchers propose that internal monitoring is part of the production system and can be achieved within the phonological encoder (Postma, 2000; Schiller et al., 2006; Wheeldon & Levelt, 1995). This is what Schiller et al. (2006) called the "production monitor," which scans phonological encoding as it builds up, as opposed to the "perception monitor," which scans the output of phonological encoding that is fed to the "general" SCS via the inner, perceptual loop. On the latter account (Levelt, 1983, 1989), the SCS would process overt speech and inner speech in the same way: this is the "perceptual loop" theory. Note that different kinds of monitoring may correspond to different loci, that is, may engage different monitoring architectures. In particular, monitoring for errors versus some phonological properties may engage different devices. As for the second issue – the nature of the code examined during internal monitoring –, it is generally accepted that this code is somehow more abstract than the code examined to monitor overt speech (Wheeldon & Levelt, 1995: see below). But its precise nature requires further qualification. In his original "Blueprint" model, Levelt (1989) viewed inner speech as a "phonetic plan," explicitly referring to the gestural scores of articulatory phonology (Browman & Goldstein, 1992). In a revised version of the model (Levelt et al., 1999), the code that is internally monitored is no more the "phonetic plan" but the metrically specified, abstract, and presumably symbolic phonological code that is elaborated just before syllabic gestural scores are retrieved or computed. This conceptual shift was motivated by findings of Wheeldon and Levelt (1995). In this seminal study, the first to use phoneme and syllable internal monitoring (hence, *not error-monitoring*), the authors found incremental serial effects in phoneme detection, reflecting either left-to-right build-up of the code within the phonological encoder or left-to-right analysis within SCS (for replications, see Morgan & Wheeldon, 2003; Schiller et al., 2006; Wheeldon & Morgan, 2002). Because the serial effect still obtained under a condition of articulatory suppression, Wheeldon and Levelt concluded that the code being monitored was more abstract than the phonetic plan, which they took as articulatorily involved (cf. Levelt et al., 1999, for detail). The model's revision has several consequences. First, the input to the inner loop in the revised model is the abstract, symbolic code mentioned above, in place of the phonetic plan. Second, a "phonetic encoder" must translate the abstract code into the phonetic plan. The new architecture of the Lemma model thus is substantially modified (see Laganaro, 2019). At the same time, Levelt et al. (1999) leave open the issue of a production vs. perception monitor. Perhaps, the two views may not exclude each other (Levelt et al., 1999; Wheeldon & Levelt, 1995), according to whether errors or, for example, phonemes are monitored for.

1.3. Monitoring for errors vs. phonological properties

Indeed, as we just noted, internally monitoring for errors might not engage the same mechanisms as monitoring for phonological properties. Current research on self-monitoring indeed suggests different mechanisms. It has been suggested that the two types of monitoring might differ in terms of conscious control (Vigliocco & Hartsuiker, 2002): error-detection would be automatic and irrepressible, whereas detecting a pre-specified property – a metalinguistic task – presumably is consciously controlled. Among the many studies conducted on self-monitoring for errors, some have addressed the issue of whether this monitoring could involve a perceptual loop, that is, explicitly examined the plausibility of the perceptual loop hypothesis for error-monitoring. Overall, a negative conclusion dominates. Hartsuiker and Kolk (2001) found support for perceptual-loop involvement in a study modelling error-to-cutoff and cutoff-to-repair time intervals: adding an inner loop feeding the SCS to their model clearly improved the fit with empirical data (but see Nooteboom & Quené, 2019, for

a critical re-evaluation of these findings). On the other hand, recent fMRI data on speech-error internal monitoring do not show speech-perception specific brain activity (e.g. Gauvin et al., 2016). In a recent review article, Gauvin and Hartsuiker (2020) propose a new model of verbal monitoring for errors, which is based on previous "conflict monitoring" models; they dismiss the perceptual loop as a plausible device for error-detection on several grounds. A strong argument, we believe, is that the perceptual loop hypothesis leaves underspecified the process of comparison between intended speech and the speech building-up, although such comparison clearly must be the basis for error-detection. The perceptual loop hypothesis, however, might be correct for the kind of "metalinguistic" monitoring best illustrated in Wheeldon and Levelt (1995; also see Van Turennout et al., 1998, for phonemic categorisation). We now turn to other than error-monitoring evidence against the perceptual loop hypothesis.

1.4. "Saturated channel" counter-evidence to perceptual loop

Let us mention briefly arguments against the perceptual loop hypothesis that do not pertain to the error-monitoring case. The underlying logic is that a single inner-loop channel for monitoring both inner and overt speech has both phenomenological and processing implications. When producing any spoken utterance, speakers should experience perceiving the overtly spoken utterance as an "echo" of the internally "spoken" utterance, assuming overt speech closely follows inner speech (by about 500 ms according to Nooteboom & Quené, 2017) in the inner loop channel. Speakers do not report such "echo" experience (Nozari et al., 2011; Vigliocco & Hartsuiker, 2002). But this might be explained by the default low attention to inner speech in the absence of specific instructions (see Sato et al., 2004). A recent study by Hansen et al. (2019), using a go-stop paradigm adapted to speech production, also argues against a single channel for overt and inner speech processing. The authors found no behavioural or brain-activity difference between two conditions: phonologically unrelated vs. overlapping go and stop stimuli (picture's name and overtly spoken word, respectively). Hansen et al. (2019) reasoned that if the inner speech for the go word and the outer speech for the stop word were feeding a single, unique perceptual loop, correct halting performance would be affected in the phonological overlap condition. It was not, suggesting separate processing routes for inner and outer speech.

1.5. Evidence for perceptual loop from phoneme or syllable internal detection

In its restricted version of a device for metalinguistic analysis, the perceptual loop gradually conveys the output of the phonological encoder to the SCS just like it conveys to it external speech. That is, the information flow analysed by the SCS is *sequential* for internal as well as external speech. A prediction derived from this proposition is that presumably perception-specific effects found with overt, external speech might be found for internal monitoring of inner speech.

Support for this version of the perceptual loop hypothesis first comes from the very clear syllabic effect found in internal syllable detection by Wheeldon and Levelt (1995: Experiment 2), using Dutch words. The effect is similar to that found for overt speech in Dutch (Zwitserlood et al., 1993). A further supporting piece of evidence is provided by Özdemir et al. (2007). They observed, in an internal phoneme detection experiment on picture names in Dutch, that monitoring latencies depend on the target phoneme's position relative to the uniqueness point (UP) of the word generated internally. (UP is the phoneme at which a word, scanned from left to right, becomes "unique" in the sense there is no other word sharing with it the speech portion up to that phoneme.) 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 (Frauenfelder et al., 1990; Marslen-Wilson, 1990; Pitt & Samuel, 1995) and makes a particularly strong case for the (restricted) perceptual-loop hypothesis. Indeed, UP effects hinge on lexical access from a phonemic-string *sequential* input: word recognition is faster for earlier occurring UP. Since lexical access has already occurred upon picture name recognition, Özdemir et al.'s results entail that lexical access is achieved "anew" a second time by the SCS from the inner speech input feeding an inner, perceptual loop.

In this paper, we focus on yet another effect, the "syllable advantage," found in overt speech perception, and test whether it could also be observed in internal speech monitoring. This test thus can be viewed as a diagnostic tool for the restricted perceptual-loop hypothesis. At the same time, the effect at stake is related to the issue of speech representation and is therefore relevant to our search of a common code for overt and inner speech. The effect 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 (Foss & Swinney, 1973; McNeill & Lindig, 1973; Savin & Baver, 1970; Swinney & Prather, 1980). Note that Swinney and Prather (1980) only found a non-significant advantage for CVC syllables over C consonants, and this when the V context was restricted to a single vowel, suggesting possible anticipation strategies. Norris and Cutler (1988) 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. (1981) used no foils and found a robust advantage for CV syllables over C stop consonants in French: /ba/ was detected faster than /b/ in *bateau* /bato/ "boat" as well as in /ba/-initial *nonwords*. Segui et al. (1981) explained the advantage of the syllable over the consonant as possibly due to the early availability in perception, at least in stop + vowel CV sequences, of *both* the consonant and the vowel at the same time (Blumstein & Stevens, 1980; Stevens & Blumstein, 1978), that is, the availability of CV as a whole, due to listeners' implicit knowledge of CV coarticulation. In other words, these authors' interpretation was that CV is an immediately available unit of perception, whereas C requires reanalysis of CV to be identified. (The influence of syllable complexity on the detection of word-initial C phonemes in overt speech also suggests that C detection follows syllable identification [Segui et al., 1990]). Could Segui et al.'s (1981) interpretation hold for inner speech, which logically does not include auditory coarticulation cues? This question seems prima facie quite challenging.

The issue we address here is thus of whether the faster monitoring of word-initial CV syllable than C consonant can also be found when monitoring internal speech. Note that it may not be so much the "syllableness" of CV that is of importance but, rather, that CV contains C. We nonetheless use the term "syllable advantage effect" for sake of simplicity. We used pictures whose name in Spanish was two- or three-syllable long and began with a CV syllable. Spanish-speaking Argentinian participants had to monitor for either the word-initial C or CV of the pictures' name, without uttering the name. The design of Experiment 1 closely followed that of Segui et al.'s (1981) study, except that participants had to monitor internally generated rather than externally presented word-forms. In Experiment 2, we tested the robustness of the syllable effect found in Experiment 1 by introducing foil items challenging syllable-monitoring more than phoneme-monitoring.

2. Experiment 1: internal monitoring of word-initial C vs. CV in picture names

2.1. Methods

Participants. Eighty-one students at the National University of Cordoba, Psychology Department, aged 18–26 years, participated voluntarily in the experiment. All were Argentinian native speakers of Spanish, with normal or corrected-to-normal vision and no known language disorder. Only 16 of the 81 participants had some basic knowledge of English. In published previous studies on phoneme or syllable detection in internal speech (Manoiloff et al., 2013, 2016; Morgan & Wheeldon, 2003; Özdemir et al., 2007; Wheeldon & Levelt, 1995; Wheeldon & Morgan, 2002), sample size varied from 20 to 40 subjects. We thus considered an 81 subjects sample-size as sufficient, at least to test for non-null effects. We nonetheless performed an a posteriori power analysis (see e.g. Althouse, 2021, questioning such practice), using G*Power 3.1 (Faul et al., 2009) for estimating an appropriate minimum sample size along the default recommendations of Cohen (1988). Because there were two measurements per subject (phoneme and syllable detection), we used the settings for a matched-pairs *t*-test with standard power 0.8 and (by subject) effect size 0.494 (see Results section). This yielded a sample size of 35 (see the full Power × Sample-size curve in Supplementary Materials).

Materials. Twenty black-on-white line drawings of simple common objects served as test picture stimuli, whose name began with a variety of Cs and CVs serving as detection targets in the experiment. They were selected from the set described in Cycowicz et al. (1997), which includes the 260 drawings in Snodgrass and Vanderwart (1980). Naming agreement for the selected experimental pictures was above 80% according to the norms established for Argentinian Spanish (Manoiloff et al., 2010). The average frequency of the 20 test picture names was of 10.8 occurrences per million (opm) (range: 0.54–48.04), according to the LEXESP Spanish lexical database by Sebastián-Gallés et al. (2000). To check whether LEXESP frequencies, which were computed for Spanish in Spain are relevant for the Argentinian subjects, we collected ratings of subjective frequencies on a 1–5 scale (5 for frequent) from 25 Argentinian native speakers of Spanish (age range 20–26 years), who did not participate in the experiments. The subjective frequencies of the 20 test picture names was 1.7 in average (range 1.1–2.5), and rather poorly correlated with the database frequencies, $r(18) = 0.342$, $p = 0.129$. Because there was more dispersion in the database than subjective frequencies (coefficients of

variation (SD/mean): 0.86 vs. 0.29 respectively), we retained the subjective frequencies to dispatch the materials evenly across two subsets (see Design section and Appendix A). There were ten word-initial target consonants (/b, d, g, p, t, k, v, f, m, n/) and thirteen word-initial target syllables (/bo, de, ga, pa, pe, pi, ti, ko, va, ve, fo, mo, ni/). Note that, in Spanish, grapheme–phoneme correspondences are largely consistent. This was the case for the word-initial letter in the names of the target pictures we used, except "c", which is pronounced /θ/ (or /s/) before "i" or "e" and /k/ otherwise. In all the test target pictures we used, "c" was followed by "o", hence pronounced /k/.

Design. The set of 20 test pictures was divided into two subsets, as balanced as possible in terms of picture name's subjective frequency, number of syllables and phonemes, and broad type of onset consonant (see Appendix A). Both subsets comprised 10 test pictures plus 82 (subset 1) or 79 (subset 2) filler pictures. The pictures of each subset were blocked by either phoneme (i.e. consonant) or syllable target, as detailed in Appendix B. Each block contained about 8 times (from 4 to 13 times) more filler than test items. There were five phoneme-target and six or seven syllable-target blocks in each subset. 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, though not for subset presentation order. We chose this option in order to keep the number of subject groups to a minimum. Had the results shown a clear group effect, we would have added another two groups to counterbalance for subset order. Each participant saw each of the 181 pictures only once, for either phoneme or syllable monitoring. Within each block, pictures were presented in a predetermined random order ensuring that test items did not appear in block-initial 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 one of three different random orders: Each participant was randomly assigned to one order.

Procedure. Participants were tested individually on phoneme- then syllable-detection or vice versa. Each phoneme- or target-block was introduced with the oral specification of the target to monitor for. In the case of the /p/-target block, for example, the phoneme-target /p/ was specified as follows: "Pulsar el botón de respuesta, lo más rápido que pueda, si y solo sí, el nombre de la imagen comienza con el sonido [pe], como en 'pera', 'papa' o 'pincel'." ["Press the response button, as quickly as possible, if and only if the name of the picture begins with the sound [pe], as in … "]. Similarly, for the /pa/-target block, the syllable-target /pa/ was specified by replacing " … comienza con el sonido [pe], como en … " [" … begins with the sound [pe], as in … "] with " … comienza con la secuencia [pa], como en 'papa', 'paleta' o 'pava'" [" … begins with the sequence [pa], as in … "]. We used the term "sequence" instead of "syllable" because participants may not have a clear and uniform notion of what a spoken syllable is. Within each block, each trial consisted of the following sequence of three or two visual stimuli displayed at the centre of the screen: the prompting message "press the spacebar to begin" (in Spanish), displayed until the participant pressed the space bar to initiate the first trial of the block; the sequence "##" displayed for 2 s as a fixation stimulus; 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. The (self-paced) experimental session typically lasted about 30 min. Participants were comfortably seated in a dimly lit quiet room. The experiment was run on a personal computer, using the DMDX experimentation software (Forster & Forster, 2003). The participants were instructed to orient their gaze at the centre 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 picture's name began with the target specified for the current series. They were instructed not to utter the picture's name. Response times were measured from the onset of picture presentation.

2.2. Results

We first computed the by-item and by-subject overall miss rate distributions. No item was discarded. One participant was a clear outlier (50% misses) in the by-subject distribution of misses (Figure 1) and was therefore excluded. For the response time (RT) data, no further filtering was applied than the 2 s time-out of the experimental set-up. (RTs ranged from 191 to 1943 ms.) The miss rate and RT data are summarised in Table 1. RTs were log-transformed to improve distribution normality (Figure 2).

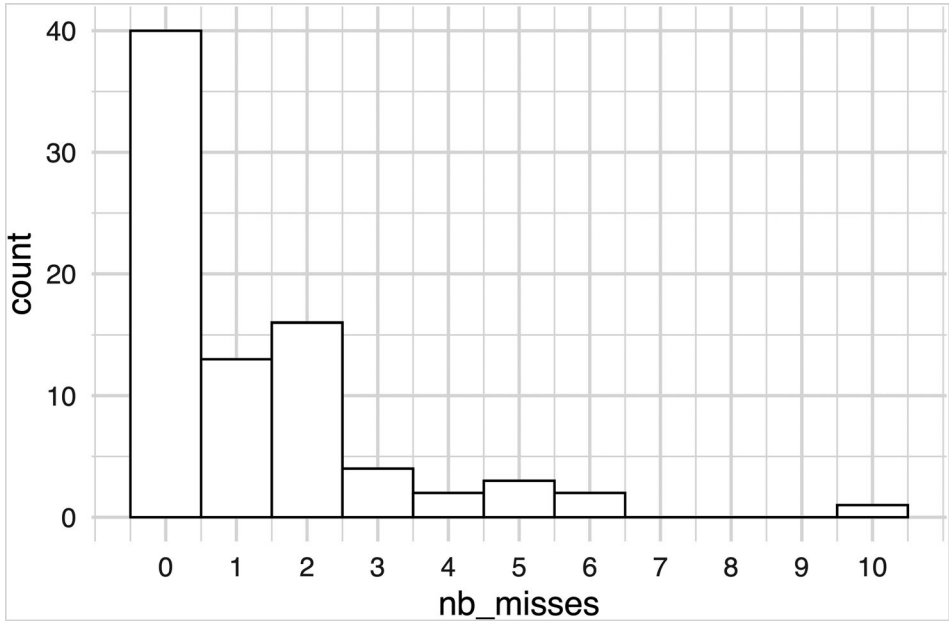

Figure 1. Experiment 1: by-subject distribution of misses.

As can be seen, syllables were detected faster and missed less often than phonemes. This was confirmed by linear mixed model analyses on the RT and miss rate data (logit regression analyses for the latter), using the *lme4* package in R (R development core team, 2016). We started with models including all the fixed effects – Order (phoneme- vs. syllable-detection first), Target (phoneme vs. syllable), and their interaction – and random subject and item intercepts as random effects. We then removed fixed effects one by one to test for their significance, using Chi-square statistics to compare models' outputs using the *anova* function. (*anova* also served to estimate model's BIC/AIC fit with the data.) We further refined the best-fitting models by attempting to remove random intercept or add random slope effects on Target and/or Order where possible: random effects leading to "singular fit" were excluded (see Supplementary materials). For miss-rates, the best-fitting models only included the random subject and item intercepts; for RTs, random subject slope on Target improved the fit. For RTs, Target was significant (shorter RTs for syllable than phoneme detection: 766 < 848 ms), $\chi^2(1) = 45.56$, $p < .0001$, and neither Order nor Order × Target were significant ($p = .585$ and $p = .631$, respectively). Cohen's d effect size for Target, computed on the by-item or by-subject data was 1.586 (large) or 0.494 (small), respectively. For miss rates (logit regression analysis), Target was significant (less misses for syllable than phoneme detection: 3.88 < 7.63%), $\chi^2(1) = 10.69$, $p = .0011$, and neither Order nor Order × Target were significant ($p = .385$ and $p = .130$, respectively). (The absence of an Order effect or an Order × Target interaction validated the choice we provisionally made not to counterbalance subset presentation order across subjects.)

Table 1. Mean RT (ms) and miss rate (%) (with SDs) in Experiment 1 according to target.

	RT	Misses
Phoneme Target	848 (183)	7.62 (12.65)
Syllable Target	766 (186)	3.88 (8.03)

Segui et al. (1981) found that response times for the detection in overt speech of CV syllables and C consonants for each item were positively correlated (by-item Pearson's *r* for 36 items, $r(34) = 0.62$, $p < .01$), suggesting that C detection depended on and *followed* CV detection. (The authors did not report correlation for miss rates.) Our internal monitoring data also reveals such a correlation (Pearson's *r*) for RTs in both the by-item data (20 items), $r(18) = 0.46$, p=.036, $CI = [0.03, 0.75]$, and the by-subject data (80 subjects), $r(78) = 0.79$, $p < .0001$, $CI = [0.69, 0.86]$. Non-parametric Spearman's ρ, however, is more appropriate to our RT or miss-rate data, which were not normally distributed. For RTs, the Spearman rank-correlations were $\rho = 0.52$, $p = .021$, and $\rho = 0.79$, $p < .0001$, for the by-item and by-subject data, respectively. For miss rates, the by-item correlation was significant, $\rho = 0.70$, $p = .0007$, but the by-subject correlation was not, $\rho = 0.15$, $p = .19$. Lack of correlation for misses in the by-subject data is likely due to the small number of cumulated misses (0–6) for each subject, as well as the large proportion of subjects (50%) with no misses for either phoneme or syllable. Nevertheless,

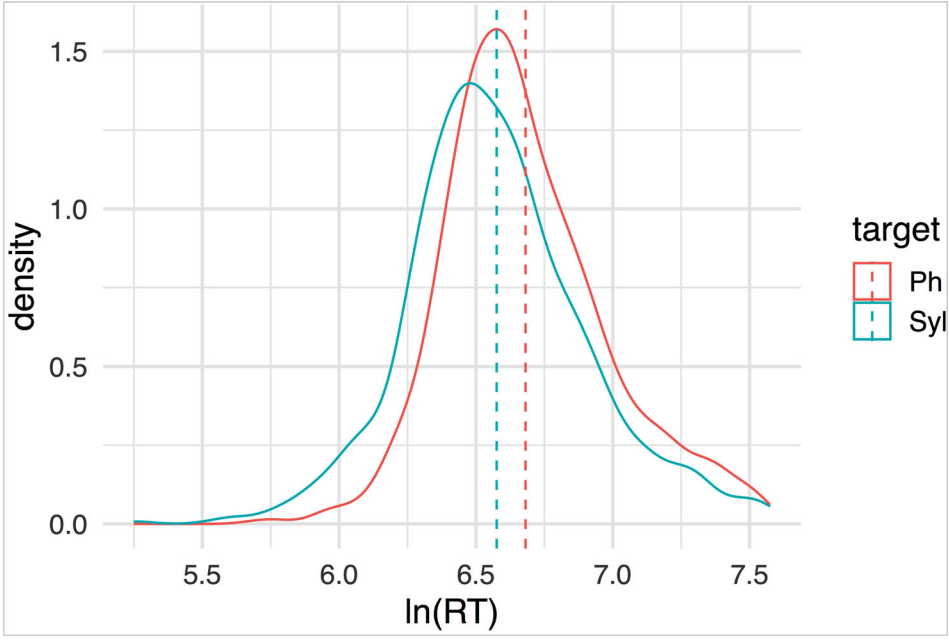

Figure 2. Experiment 1: density distributions of ln(RT) for phoneme- and syllable-detection.

focusing on the comparable correlation data in Segui et al.'s (1981) study and ours, the two sets of data are quite congruent, thus suggesting that internal C-detection depends on internal CV-detection.

2.3. Discussion

Compared to the response times usually found for overt speech, the RTs we report for inner speech are much longer: about 800 ms in average, compared to, for example, around 300 ms in Segui et al. (1981) or around 500 ms in Norris and Cutler (1988) in their no-foil conditions. Most relevant to our study, Morgan and Wheeldon (2003) and Wheeldon and Morgan (2002) compared detection RTs in inner and overt speech: they found longer RTs for inner than overt speech by ~550 ms in average. Because target-detection RTs are measured from target's time of occurrence in overt speech but from prompt presentation in inner speech, the longer RTs for inner than overt speech are expected in Levelt's model since they include the amount of time necessary to generate the "phonetic plan." Indefrey and Levelt (2004; also see Indefrey, 2011) estimated this additional time to 600 ms from picture presentation on the basis of brain activity data.

A possible concern with our results is that phoneme detection might have been challenging 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 Swinney and Prather (1980), consonant detection is slowed down by the "uncertainty as to the identity of the vowel following the target consonant" (Swinney & Prather, 1980, p. 104). 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 slow-down of phoneme detection latencies. To sum up, the materials of Experiment 1 might have disfavoured, 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 disfavoured by the presence of CV' foils in CV-target blocks.

In their study, Segui et al. (1981) did not use foils (i.e. catch trials) to encourage the participants to use a cautious strategy and fully analyse and check the speech stimuli (words or nonwords) for their phonetic content. This allowed them to explore a pre-lexical level of processing and, indeed, they obtained the same results for words and nonwords, with very short RTs for both (about 280 vs. 345 ms for syllable vs. phoneme

detection). Norris and Cutler (1988), in their "no-foil" conditions, found a reversed pattern for words – though not for nonwords (see their Table 2) – but with much longer RTs. A possible reason for this discrepancy could be the CVC format of the target syllables used, a possibly post-lexical level of processing, or that English was used rather than French (but see Foss & Swinney, 1973; McNeill & Lindig, 1973; Savin & Baver, 1970).

3. Experiment 2: control experiment with CV' foils for CV targets

Thirty CV' foils were added to the twenty CV target-bearing test items in the syllable-target blocks. That is, 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 Norris and Cutler's (1988) findings, this should slow down CV detection substantially.

3.1. Methods

Participants. Fifty-two students at the National University of Cordoba, Psychology Department, aged 18–26 years, participated voluntarily in the experiment. All were Argentinian native speakers of Spanish, with normal or corrected-to-normal vision and no known language disorder. None of them had participated in Experiment 1. An a posteriori power analysis (see Experiment 1's *Participants* section) indicated that a 52 sample-size only ensured a 0.65 power, given the effect size found (d = 0.321). But note this mainly indicates that our 52-subjects sample size does not warrant avoiding type II errors (accepting null hypothesis when it is false).

Materials and design. The same twenty test pictures as in Experiment 1 were dispatched into the same two subsets as in Experiment 1 (Appendix A). They were associated with the same 10 phoneme or 13 syllable targets as in Experiment 1. The only difference with Experiment 1 was the addition of 30 "foil pictures" (henceforth, foils) whose name was beginning with CV' in the 13 CV syllable-target blocks (15 foils in each subset). No foils were added to the phoneme-target blocks. Appendix C shows the distribution of the 30 foils in the 13 syllable-target blocks. As in Experiment 1, the experimental design was counterbalanced across subjects for target-type order: one group detected phonemes then syllables and the other group syllables then phonemes in subsets 1 and 2.

Procedure. The procedure was the same as in Experiment 1.

Table 2. Mean RT (ms) and miss rate (%) (with SDs) in Experiment 2 according to target.

	RT	Misses
Phoneme Target	948 (96)	12.69 (11.57)
Syllable Target	912 (94)	7.88 (9.15)

3.2. Results

We first computed the overall miss rates by item and by participant (Figure 3). We applied the same data exclusion strategy as for Experiment 1. No item or subjects was discarded. As in Experiment 1, all the available RT data were retained (RTs ranged from 530 to 1964 ms). RTs were log-transformed to improve distribution normality (Figure 4). The miss rate and RT data are summarised in Table 2. Figure 5 shows the miss rate and RT data for Experiments 1 and 2.

Syllables were still detected faster and missed less often than phonemes, although the effect was numerically weaker than in Experiment 1 for RTs. Figure 5 summarises the RT and miss rate data for Experiments 1 and 2.

We ran linear mixed model analyses on the Experiment 2 data, following the same strategy as for Experiment 1, with the same fixed effects: Target, Order, and Target × Order interaction. For either miss-rates or RTs, adding random slopes did not improve the models in terms of AIC or BIC. For RTs, Target was significant, $\chi^2(1) = 4.97$, $p = .026$: RTs were shorter for syllable than phoneme detection (912 < 948 ms). The effect size for Target, computed by-tem or by-subject was 0.244 or 0.321 (small), respectively, indicating a weaker effect than in Experiment 1. Neither Order nor Order × Target were significant, ($p = .328$ and $p = .818$, respectively). For miss-rates (logit regression analysis), Target was significant, $\chi^2(1) = 8.27$, $p = .0040$: miss rate was lower for syllable than phoneme detection (7.88 < 12.69%). Neither Order nor Order × Target were significant ($p = .137$ and $p = .389$, respectively).

As in Experiment 1, there were significant correlations between syllable and phoneme monitoring data for both RTs and miss-rates, except for by-subject miss-rates. Pearson correlations for comparison with Segui et al. (1981) were: $r(18) = 0.81$, $p < .0001$, $CI = [0.57, 0.92]$ and $r(50) = 0.35$, $p = .011$, $CI = [0.08, 0.57]$ (by-item and by-subject data). The corresponding Spearman correlations were $\rho = 0.69$, $p = .0011$ (by-item), and $\rho = 0.29$, $p = .036$ (by-subject). For miss-rates, only the by-item correlation was significant, $\rho = 0.83$, $p < .0001$. Overall, then, Experiment 2's data confirm the correlation between syllable and phoneme detection found in Experiment 1 (inner speech), and in Segui et al.'s (1981) (overt speech).

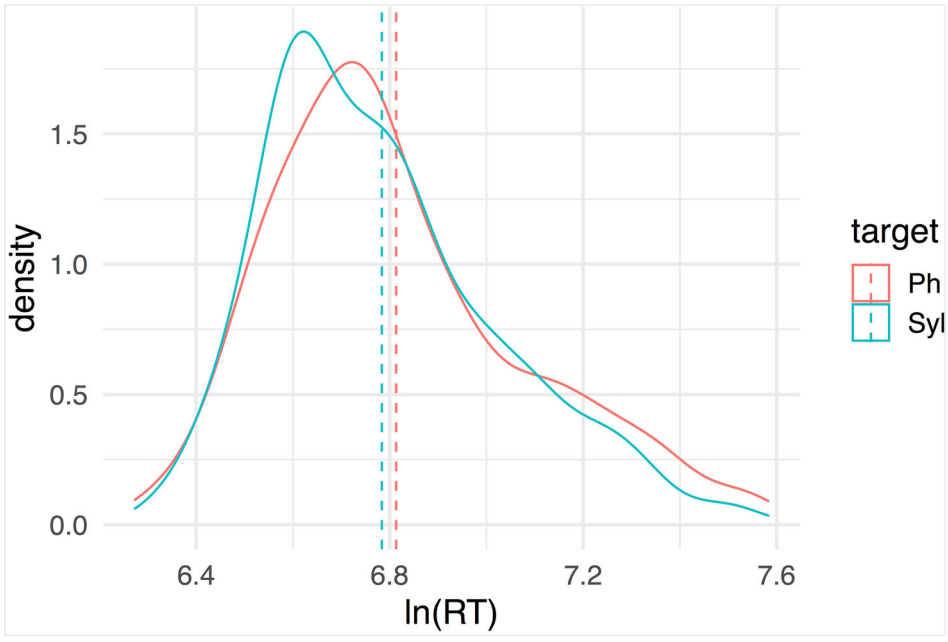

Figure 4. Experiment 2: density distributions of ln(RT) for phoneme- and syllable-detection.

The differences between the two experiments, as shown in Figure 5, called for further statistical analysis bearing on both experiments, the main question being whether the Target effects we found differed in strength between the two experiments. We ran linear mixed model analyses on the combined data of Experiments 1 and 2, following the analysis strategy described for Experiment 1, with the additional Experiment fixed effect (Experiment 1 vs. 2). For miss-rates, only random intercepts were retained: adding random slopes for items or subjects did not improve the models (BIC and/or AIC). For RTs, adding random slope on Experiment for items and for subjects improved the models in terms of BIC and/or AIC. For both RTs and miss-rates

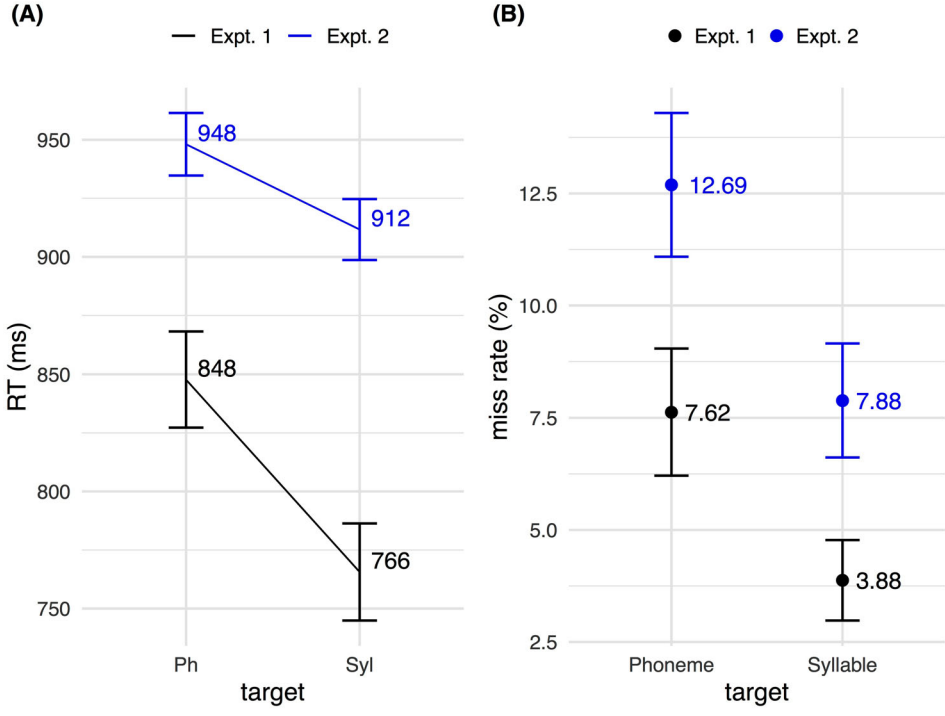

Figure 5. Experiments 1–2: by-subject (a) RT, and (b) miss rate data (error bars: ±1 SE).

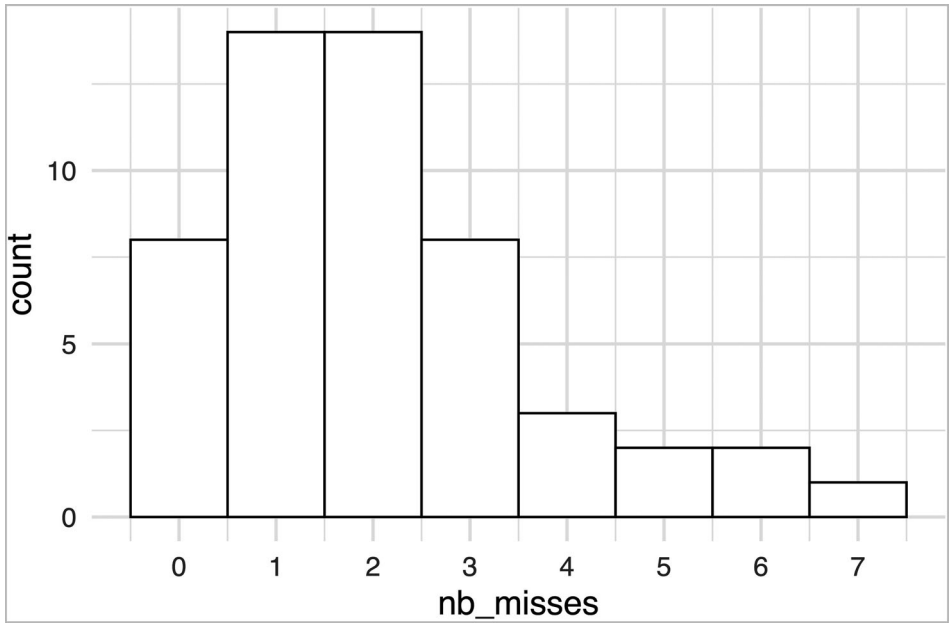

Figure 3. Experiment 2: by-subject distribution of misses.

(logit regression analysis for the latter), Target was significant, indicating faster RTs, $\chi^2(1) = 75.08$, $p < .0001$, and lower miss-rates, $\chi^2(1) = 19.47$, $p < .0001$, for syllable- than phoneme-detection across experiments. Experiment was significant too, $ps < .0005$, reflecting shorter RTs and lower miss rates overall in Experiment 1 than 2 (807 < 930 ms; 5.75 < 10.29% misses). The Target × Experiment interaction was significant for RTs, $\chi^2(1) = 17.02$, $p < .0001$, but not for miss rates, $p = .53$: the Target effect was stronger in Experiment 1 than Experiment 2 for RTs (82 vs. 36 ms) but not for miss-rates (3.74 vs. 4.81%). No other effect or interaction was found significant.

3.2. Discussion

As expected, the addition of CV′ foils in Experiment 2 slowed down syllable-detection, due to a more cautious monitoring strategy. However, the syllable advantage maintained, though substantially weakened for RTs, as shown by the significant Experiment × Target interaction. Although the phoneme-detection blocks were identical in experiments 1 and 2, phoneme-detection was much slower in the latter. In particular, the phoneme-then-syllable subject groups of each experiment received an identical *initial* block for phoneme detection. Yet, the Experiment 2 group was slower than the Experiment 1 group by some 100 ms. This cannot reflect the influence of a yet-to-come syllable-detection block with CV′ foils inducing a cautious strategy across the board. Such a "strategy contamination" could be possible, however, for the syllable-then-phoneme group, resulting in slower RTs to phoneme- compared to syllable-detection in this group than in the phoneme-then-syllable group, hence in a larger syllable advantage effect, possibly diagnosed by a Target × Order interaction. But no such interaction was observed. We therefore must conclude that the subjects in Experiment 2 were slower overall than those in Experiment 1 for some unknown reason. We however analysed the data in more detail, looking at the first trials of phoneme-detection, especially in Experiment 2. We indeed found clearly longer RTs in the first two trials than the others in the syllable-first but not the phoneme-first order of Experiment 2. This can be taken as a short-lived contamination effect from the initial block of syllable-detection with catch-trials (see Supplementary materials for more detail: "time-course_RT-analyses").

To sum up, the subjects in Experiment 2 were slower overall than those in Experiment 1 but still showed a significant syllable advantage in monitoring inner-speech. The effect was much weaker than in Experiment 1 for RTs, as shown by the Experiment × Target interaction, but remained as strong in Experiment 2 as Experiment 1 for miss rates.

4. General discussion

In the two experiments reported in this study, participants had to detect word-initial C consonant or CV syllable in the names of visually presented pictures, without uttering the names. We found in both experiments that internally monitoring for CV syllables was faster than for

the corresponding C phonemes. In Experiment 1, this syllable advantage was very clear-cut, similar to the syllable advantage found in monitoring for external speech in Segui et al.'s (1981) study (82 vs. 63 ms effects). In Experiment 2, we introduced a substantial number of CV' foil items for CV syllable detection (foil to test ratio 3:2). That is, we added pictures whose name was beginning with CV' in the CV-target blocks: the same consonant as the target followed by a different vowel. In external, overt speech, this manipulation is known to create a "phonemic garden path" for syllable monitoring and generally induces a cautious strategy. The presence of foils requires a more careful, complete phonetic analysis of the auditory input and results in much slower response times (Norris & Cutler, 1988). Concerning inner speech, there is no previous study, to our knowledge, suggesting similar foil effects. But the results of Experiment 2 clearly show that foil effects do apply to internally monitoring for inner speech as well as overt speech. Indeed, only syllable-, not phoneme-detection was affected by the CV' foils, resulting in a substantially reduced syllable over phoneme advantage. This is best explained by CV' foils slowing down CV detection, a typical foil effect.

4.1. Production or perception monitor?

Overall, then, we found a CV syllable over C consonant advantage in monitoring for inner speech, which is quite similar to that previously reported for overt speech, including foil effects. This result may be taken as supporting the restricted version of the perceptual-loop hypothesis proposed in the introduction (Schiller et al.'s [2006] perception monitor account limited to "metalinguistic" monitoring), in line with a few previous studies on internal monitoring for phonemes or syllables (Morgan & Wheeldon, 2003; Özdemir et al., 2007; Wheeldon & Morgan, 2002). Wheeldon and Levelt (1995) also found a striking parallelism between the processing of inner and overt speech in the form of "a textbook interaction of syllable target and carrier word syllabification" (Wheeldon & Levelt, 1995, p. 325). Indeed, the same syllable-match effect (cf. the seminal fragment detection study by Mehler et al., 1981) obtained in the same language (Dutch) for inner speech (Wheeldon & Levelt, 1995) and overt speech (Zwitserlood et al., 1993), though the effect was less clear-cut in the latter (see Cutler, 1997; Vroomen & de Gelder, 1994). However, Wheeldon and Levelt (1995) interpreted the syllable-match effect as reflecting the structural properties of the speech code elaborated for production by the phonological encoder (i.e. this code must be syllabified) rather than the processing strategy of a perception monitor. In a later study directly comparing fragment detection in inner and overt speech for English, Morgan and Wheeldon (2003) obtained clearly parallel results in inner and overt speech (syllable-match effect for CVC targets but not CV targets), which they interpreted as reflecting features "of a perception-based monitoring system" (Morgan & Wheeldon, 2003, p. 291). Özdemir et al. (2007) is certainly an even more solid piece of evidence suggesting that inner and overt speech are processed by a common speech perception/comprehension system (see Introduction). Moreover, the UP effect they found in inner speech entails that inner speech is treated by the speech perception system as an as yet unprocessed, fresh input on which lexical access remains to be performed. Our data also speak for a perception monitor common to both outer and inner speech. Moreover, the specific common effect at stake – the syllable advantage over phoneme – might help understanding the commonalities in the involved representations and their processing. We will turn to this representational issue later on.

Let us consider the production-monitor alternative explanation of our results in the framework of Levelt's Lemma model. In either the original 1989 "Blueprint" or the 1999 revised model, the production monitor would operate *within* the phonological [and phonetic] encoder module, either at any level of representation, as proposed by Postma (2000), or at some specific level of representation. (Note that, in spite of the difference in architecture between the original and revised Lemma models, the revised model's abstract, symbolic representation that feeds the inner loop in place of the phonetic plan must be now the output of a more narrowly specialised phonological encoder and the input to a separate phonetic encoder. This representation thus still escapes the production-monitor scope.) First, whatever the monitored level of module-internal representation, the syllable advantage cannot be easily explained within the production-monitor account. If all the levels are monitored, phonemes should be available first upon lexical access, before (re)syllabification has taken place. If some syllabified symbolic code is monitored, an analogy with print – both print and inner-speech would be sequences of phoneme-sized symbols – does not clearly suggest that syllable-sized letter sequences should be detected more easily than single letters. To our best knowledge, relevant data on fragment detection in print would rather suggest that a sequence of letters in a printed word would be detected more slowly than a single letter (Colé et al., 1999). Second, in Levelt's modularist model, the phonological encoder component is viewed as a processing module (Levelt, 1989, p. 15) as defined by Fodor (1983,

1985), that is, an *automatic* module impenetrable to conscious inspection and control ("informational encapsulation"). The production-monitor account precisely entails that inner representations *within* the phonological encoder module be monitored, thus violating "impenetrability." Moreover, recall that while monitoring for errors may be automated and may well happen within the encoder module as part of the module's missions, monitoring for inner speech properties is a metalinguistic task (Vigliocco & Hartsuiker, 2002), which must be conscious and thus is less likely to operate on module-internal representations. Therefore, were the phonological encoder a module in the classic sense, the production-monitor account seems inappropriate.

To sum up, although the production monitor account, or some hybrid account combining a perception monitor with one or several production monitoring devices cannot be dismissed, the current set of relevant data and logical arguments lead us to favour a perception monitor account, at least for the type of monitoring at stake: phoneme or syllable detection. That a production monitor may handle error-monitoring, a presumably more automatic process, is possible but this issue is out of the scope of the present study. We turn now to the issue of the common code that must be processed by the speech perception and comprehension system for both overt and inner speech.

4.2. Phonological or phonetic code?

That both inner and overt speech be monitored via a common speech perception system calls for the further assumption that they are coded with similar, compatible representations. This is at least what should be expected for the representations on which phoneme and syllable monitoring operate. We are therefore faced with a dilemma between the two versions of the Lemma model: abstract, symbolic though syllabified phonological representations in the 1999 revised version versus syllabic gestural scores making up the "phonetic plan" in the 1989 original version. The gestural scores of the phonetic plan are taken by Levelt from articulatory phonology: they describe *speech tasks* in terms of constriction gestures and timing coordination. In orthodox articulatory phonology, though, the primary specifications rather are "coupling graphs," specifying how the "coupling oscillators" for gestures are phased with each other (Goldstein et al., 2006; Goldstein et al., 2009; Nam et al., 2009). For the moment, however, we follow Levelt's formulation for sake of clarity. As we noted earlier in the Introduction section, a further motor programme elaboration stage is needed to compute how these speech tasks are to be carried out in terms of neuromuscular commands determining the eventually achieved articulatory movements that produce overt speech. As articulatory phonology would agree, we take the phonetic plan's gestural scores as *abstract*, *discrete* phonological specifications, which, however, differ profoundly from classic phonological specifications in that they specify *gestural timing* relationships. With this important qualification in mind, let us consider how the symbolic and gestural *phonological* codes compare in unifying the inner- and overt-speech codes.

First, a number of studies show that the processing of overt speech, in particular spoken word recognition but also speech segmentation and phonemic categorisation or detection, is sensitive to "phonetic detail," that is, is "phonetically" informed. Whereas some relevant findings reflect artificial effects of sub-categorical mismatch (Marslen-Wilson & Warren, 1994; Streeter & Nigro, 1979; Whalen, 1984, 1991), many intriguing findings bearing on natural speech show the effects of sometime subtle durational differences (Dufour & Meynadier, 2019; Mitterer & McQueen, 2009; Shatzman & McQueen, 2006; Snoeren et al., 2008; Spinelli et al., 2003; Spinelli et al., 2007). These findings suggest that overt speech processing relies on representations containing *timing* information. Such representations would be in line with articulatory phonology. As an illustration, voice-assimilated *soute* ("hold": underlying /sut/ > /sud/) and non-assimilated *soude* ("soda": /sud/) both surface as /sud/ ([sud] in broad IPA transcription) but differ in that [u] vowel and [d] closure durations remain typical of the underlying form (shorter [u] and longer [d] for *soute* than *soude*). Listeners are sensitive to this durational difference since only voice-assimilated *soute* primes semantically related *bagage* ("luggage"): *soude* does not prime *bagage* at all (Snoeren et al., 2008).

Second, with regards to inner speech, it seems prima facie difficult to find evidence for the role of phonetic detail in the absence of an acoustic output. However, assuming that the kind of inner speech elicited by print is similar to the kind of inner speech elicited by pictures, there is some indication that it is "phonetically informed" in much the same way as in the illustration from Snoeren et al. (2008). That is, inner speech elicited by print contains *timing* information. Now, how does it compare with the inner speech elicited by pictures to be silently, internally named? It is widely accepted that a printed sequence automatically "activates" a phonological representation (among many references, some pioneer studies: Ferrand & Grainger, 1992; Lukatela & Turvey, 1994a, 1994b; Perfetti et al., 1988; Van Orden, 1987; Ziegler & Jacobs, 1995), including subphonemic features (Lukatela et al., 2001). An interesting early account for the irrepressible phonological involvement

in reading was offered by Huey (1908/1968) who remarked that "the inner saying or hearing of what is read seems to be the core of ordinary reading" (Huey, 1908/1968, p. 122). Abramson and Goldinger (1997) argued that the "inner speech" occurring during reading may not be more abstract than overt speech in that certain well documented subphonemic phonetic regularities may affect the reading process. They reported, in particular, that phonetically longer words are recognised in print more slowly than shorter words. For example, *plead* is longer than *pleat*, due to its longer /i/ vowel. (Increase in vowel duration before voiced obstruents is only partially compensated by a decrease in the obstruent duration [e.g. Port, 1981]). Lukatela et al. (2004) replicated these findings. They reinterpreted inner-speech in reading as a "phonetically informed" phonological representation, which is better understood in terms of articulatory phonology (Browman & Goldstein, 1992, 2000) than of classic linear or autosegmental phonology. This is because articulatory phonology representations implicitly specify phonetic aspects, especially timing relationships, hence durations. In this particular case, Browman and Goldstein (1992, p. 170) interpret the cross-linguistically typical VC duration patterns for voiced vs. voiceless C as due to "overlap differences between consonant and vowel gestures" (for more detail, see Fujimura, 1981).

To sum up, a common code akin to articulatory phonology representations, in line with the "common currency" shared by speakers and listeners proposed by Goldstein and Fowler (2003), is probably a better candidate for the unifying code we postulate for both inner- and overt-speech than the classic linear, symbolic phonological code proposed in Levelt et al. (1999) revised model. Recall that their motivation for this shift from the initial "phonetic plan" proposal was that articulatory suppression did not affect the incremental effect they found in phoneme detection (Wheeldon & Levelt, 1995), pointing to a phonological rather than phonetic nature of the code being monitored (see Introduction, 1.2). However, in our current understanding of articulatory phonology, the gestural scores of the "phonetic plan" initially posited in Levelt (1989) can be viewed as phonological and abstract, and thus may well be immune to articulatory suppression. In our view, articulatory suppression can introduce perturbation at the output of the motor plan elaboration stage, a later stage within the extended phonetic encoder described by Laganaro (2019, Figure 2). To clarify, we propose the Lemma model could be revised along these lines, somehow going back to the original 1989 Blueprint model, as illustrated in Figure 6. For sake of simplicity, we retained segmental representations in the earliest processing steps but a more radical version of the diagram would abandon them altogether. For the same reason, we still use in this diagram the "classic" gestural scores, although, coupling graphs can be viewed as more basic "common-currency" specifications of speech acts for both production and perception (Goldstein et al., 2006; Goldstein et al., 2009; Nam et al., 2009).

Further qualifying these articulatory phonology constructs and examining their mental or neural implementation goes beyond the scope of the present study. We only cursorily comment on this however important aspect. One serious possibility, suggested by recent theoretical speculation as well as brain imagery data, is that listeners and speakers conduct similar (or isomorphic) prediction-driven internal *motor simulations* that yield estimations of perceptual consequences in terms of somatosensory and auditory percept (Tian & Poeppel, 2012; Tian et al., 2016; for a wider domain for these concepts, see Feldman Barrett, 2017, chapter 4: *The origin of feeling*). The common currency for speech acts advocated by articulatory phonology could thus be, concretely, motor simulation. As suggested by one reviewer, motor simulations could also allow for comparing internal or external speech with the targets to be detected. On this account, our study would illustrate a manifold common currency for inner (production) and outer (perception) speech as well as for the speech-target representation required by the task demand. We return to the latter aspect in the next section, keeping in mind that speech acts need be specified along both the vowel and consonant tiers, making an hypothetical motor simulation for a consonant incomplete.

4.3. The syllable over phoneme advantage

Would our assumption – that the code for inner speech feeding the inner loop is akin to articulatory phonology representations – help understand the syllable advantage we found in internal monitoring? A crucial, basic tenet of articulatory phonology is gestural phasing relationships (Browman & Goldstein, 1989, 1990, 1992, 1995; Goldstein et al., 2006). In an utterance-initial CV sequence, C and V are coproduced "in phase" (Browman & Goldstein, 1988, 1992, 2000; Goldstein et al., 2009; Nam et al., 2009), a notion originally proposed by Öhman (1966) and Fowler (1980). Put another way, CV is a *cohesive unit in production* (the coupled oscillators for the associated gestures are tightly time-locked). Assuming that perceiving speech is directly perceiving speech acts in the form of gestures and inter-gestural timing, CV should be a *cohesive, basic "unit" in perception* as well. Hence, whereas identifying

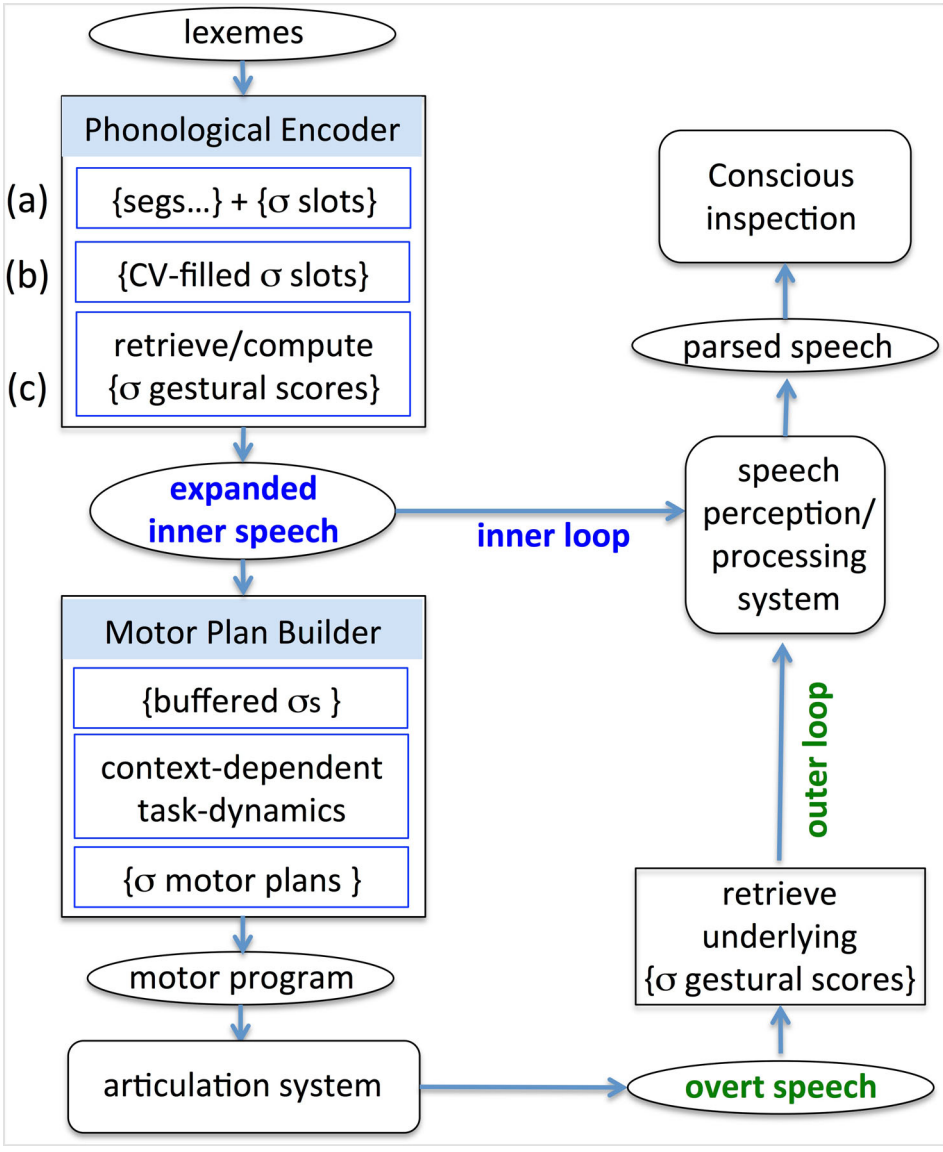

Figure 6. Functional diagram for a revised Lemma model; the aspects relevant to error monitoring are not addressed. Curly brackets enclose representations; σ stands for syllable; (a-c) stand for the initial, intermediate, and final representations of the phonological encoder. The motor plan elaboration stage ("builder") is speculative.

and detecting CV is primary, C or V are "recoverable" (identified and detected) only in a second step of (conscious) analysis. This account would explain the CV over C advantage that is found in either overt or inner speech, assuming that both are coded in terms of the common currency posited by articulatory phonology. Under this assumption, CV is a cohesive unit, with tightly time-locked C and V gestures; its analysis yields C and V as later available sub-units.

Segui et al. (1981) interpreted the CV over C advantage (C being a stop) in a rather similar way in that they proposed that C and V are perceived *at the same time*, as parts of the whole unit CV, based on Blumstein and Stevens' (1980) finding that invariant *acoustic* cues in the beginning of a word-initial CV (Stevens & Blumstein, 1978) convey the perception of CV *as a whole*. A very short portion of this beginning (about 40 ms: stop release burst and beginning of the transition from C to V) was found sufficient to convey the cohesive percept. In the case of inner speech, however, acoustic information is obviously not available. The account proposed by Segui et al. (1981) therefore cannot apply to inner speech. But our proposal that the cohesive percept of a CV structure is due to the tight in-phase coordination of the C and V gestures can apply to both overt and inner speech, thus preserving the basic idea of Segui et al. (1981) that CV is a primary unit of perception. The relatively invariant acoustic pattern at the very beginning of CV invoked by Segui et al. may be viewed as a mere *consequence* of the tightly time-

locked gestural specification. To sum up, the advantage of the articulatory phonology over the CV acoustic invariance account is that it explains the CV over C advantage for both overt and inner speech. On another articulatory-phonology-related account, both the speech targets (C or CV) and the inner or outer speech to be monitored for (CV words) would be represented as motor simulations in subjects' minds. As we noted, the motor simulation for CV is complete or self-sufficient. That for C must ignore the vowel-tier gestural specification (and motor simulation) and is therefore conceivably more difficult to compare with the inner or outer speech utterance motor simulation, which includes the vowel tier. This account is structurally similar to the one we propose but may be more realistic in terms of the actual inner workings of speech-fragment monitoring.

Note that our proposed accounts only apply to CV structures. This is one limitation of our study. How would the articulatory phonology account fare with more complex structures such as CVC syllables? This is of course an empirical question. For CVC syllables, the coda consonant is viewed as produced in anti-phase relative to onset C or to V. That is, the phasing relationship of coda C with syllable-initial CV is sequential, hence less stable than that for the CV structure (Browman & Goldstein, 1988; Goldstein et al., 2006; Nam et al., 2009). On this view, CVC must be a less cohesive unit in production and perception than CV. As a possible consequence, monitoring for CVC syllables might be less easy and less straightforward than monitoring for cohesive CV structures. We predict that CV is easier to detect than CVC due to (i) the greater CV cohesiveness and (ii) the relative autonomy of the C coda: in articulatory phonology terms, CVC does not make a tightly time-locked "molecule." The monitoring performance for such a structure may therefore be more sensitive to metalinguistic strategies. This seems to be the case for overt speech: CVC detection can be faster than C detection when CVC is predictable from CV (Foss & Swinney, 1973; McNeill & Lindig, 1973; Savin & Baver, 1970) but becomes clearly slower than C detection when predictability is cancelled out by the presence of CVC' foils (Norris & Cutler, 1988). CV structures therefore might be better cohesive units than CVC linguistic syllables for coding the speech input to the comprehension system, be it overt speech or inner speech. In their discussion of the "mental syllabary" accessed to during speech production, Levelt and Wheeldon (1994) consider an alternative to syllabic gestural scores along the lines of the demisyllables proposed by Fujimura (1990). On the other hand, recent articulatory phonology advances give more weight to the syllable construct, proposing to integrate into the basic coupled-oscillators organisation the syllabic jaw oscillation proposed by MacNeilage's (1998) frame/content theory (Goldstein et al., 2006, p. 239) (also see Montani, Chanoine, Grainger, & Ziegler (2019) for recent evidence of syllable-specific neural activation in reading and prepared speech production). Further research is needed to pursue this representational issue in both overt and inner speech. Again, parallel findings between overt and inner speech would be a diagnostic for testing the restricted perceptual-loop theory: Do we analyse overt and inner speech in the same way, that is, by relying on the same representations and processes?

To conclude, the present study may be viewed as one step forward on this research avenue. Adding to previous findings (Morgan & Wheeldon, 2003; Özdemir et al., 2007; Wheeldon & Levelt, 1995), the data we report support the view that we listen to inner speech just like we listen to outer speech when we monitor pre-articulatory speech production for phonological properties. That is, our study supports a "restricted" version of the perceptual-loop hypothesis

Several limitations, however, need be acknowledged. One is inherent to the experimental paradigm: inner-speech elicitation using pictures. The limitation is in the number of picture-names receiving sufficiently high agreement and allowing for the critical experimental comparison. Another limitation is the possibly language-specific significance of the results. Other languages, in particular belonging to various rhythmic classes, need be investigated in order to tease apart language-specific and universal aspects of overt- and inner-speech processing similarities (e.g. stress-timed English, mora-timed Japanese, etc.).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by a public grant from the French National Research Agency (ANR) as part of the programme "Investissements d'Avenir" (reference: ANR-10-LABX-0083) to the 1st, 3rd, and 4th authors, and a Secyt (Secretaria de Ciencia y Técnica) grant from Universidad Nacional de Córdoba (Argentina).

Data availability statement

Supplementary Materials comprising full description of the stimulus materials, raw result data, scripts to process the raw data, R-analysis scripts (plain and mark-down formats) and outcomes (html files), as well as additional figures and tables, are publicly available at: https://osf.io/92ewg/ (OSF repository,

"syllable_vs_phoneme" project); https://doi.org/10.17605/osf.io/92ewg

ORCID

Pierre Hallé http://orcid.org/0000-0002-3600-2070

References

Abramson, M., & Goldinger, S. (1997). What the reader's eye tells the mind's ear: Silent reading activates inner speech. *Perception and Psychophysics*, *59*(7), 1059–1068. https://doi.org/10.3758/BF03205520

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931–965. https://doi.org/10.1037/bul0000021

Althouse, A. D. (2021). Post Hoc power: Not empowering, just misleading. *Journal of Surgical Research*, *259*, A3–A6. https://doi.org/10.1016/j.jss.2019.10.049

Basho, S., Palmer, E., Rubio, M., Wulfeck, B., & Müller, R.-A. (2007). Effects of generation mode in fMRI adaptations of semantic fluency: Paced production and overt speech. *Neuropsychologia*, *45*(8), 1697–1706. https://doi.org/10.1016/j.neuropsychologia.2007.01.007

Blumstein, S., & Stevens, K. (1980). Perceptual invariance and onset spectra for stop consonants in different vowel environments. *The Journal of the Acoustical Society of America*, *67*(2), 648–662. https://doi.org/10.1121/1.383890

Browman, C., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, *45*(2-4), 140–155. https://doi.org/10.1159/000261823

Browman, C., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251. https://doi.org/10.1017/S0952675700001019

Browman, C., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, *18*(3), 299–320. https://doi.org/10.1016/S0095-4470(19)30376-6

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180. https://doi.org/10.1159/000261913

Browman, C., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 175–193). The MIT Press.

Browman, C., & Goldstein, L. (2000). Competing constraints on intergestural coordination and self-organization of phonological structures. *Les cahiers de l'ICP*, *5*, 25–34.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Lawrence Erlbaum Associates.

Colé, P., Magnan, A., & Grainger, J. (1999). Syllable-sized units in visual word recognition: Evidence from skilled and beginning readers of French. *Applied Psycholinguistics*, *20*(4), 507–532. https://doi.org/10.1017/S0142716499004038

Corley, M., Brocklehurst, P., & Moat, H. (2011). Error biases in inner and overt speech: Evidence from tongue twisters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 162–175. https://doi.org/10.1037/a0021321

Cutler, A. (1997). The comparative perspective on spoken-language processing. *Speech Communication*, *21*(1-2), 3–15. https://doi.org/10.1016/S0167-6393(96)00075-1

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*(2), 171–237. https://doi.org/10.1006/jecp.1996.2356

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124–142. https://doi.org/10.1016/0749-596X(88)90070-8

Dufour, S., & Meynadier, Y. (2019). Temporary ambiguity in whispered word recognition: A semantic priming study. *Journal of Cognitive Psychology*, *31*(2), 157–174. https://doi.org/10.1080/20445911.2019.1573243

Faaborg-Andersen, K., Edfeldt, Å, & Nykøbing, F. (1958). Electromyography of intrinsic and extrinsic laryngeal muscles during silent speech: Correlation with reading activity: Preliminary report. *Acta Oto-Laryngologica*, *49*(1), 478–482. https://doi.org/10.3109/00016485809134778

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Feldman Barrett, L. (2017). *How emotions are made*. MacMillan.

Fernyhough, C. (2004). Alien voices and inner dialogue: Towards a developmental account of auditory verbal hallucinations. *New Ideas in Psychology*, *22*(1), 49–68. https://doi.org/10.1016/j.newideapsych.2004.09.001

Ferrand, L., & Grainger, J. (1992). Phonology and orthography in visual word recognition: Evidence from masked nonword priming. *The Quarterly Journal of Experimental Psychology Section A*, *45*(3), 353–372. https://doi.org/10.1080/02724989208250619

Fodor, J. (1983). *The modularity of mind*. MIT Press. https://doi.org/10.7551/mitpress/4737.001.0001

Fodor, J. (1985). Precis of the modularity of mind. *Behavioral and Brain Sciences*, *8*(1), 1–42. https://doi.org/10.1017/S0140525X0001921X

Forster, K., & Forster, J. (2003). DMDX: A windows display program with millisecond accuracy. *Behaviour Research Methods, Instruments, and Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

Foss, D., & Swinney, D. (1973). On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, *12*(3), 246–257. https://doi.org/10.1016/S0022-5371(73)80069-6

Fowler, C. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133. https://doi.org/10.1016/S0095-4470(19)31446-9

Frauenfelder, U., Segui, J., & Dijkstra, T. (1990). Lexical effects in phonemic processing: Facilitatory or inhibitory? *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 77–91. https://doi.org/10.1037/0096-1523.16.1.77

Fujimura, O. (1981). Elementary gestures and temporal organization—what does an articulatory constraint mean? In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive*

representation of speech (pp. 101–110). North-Holland. https://doi.org/10.1016/S0166-4115(08)60182-X

Fujimura, O. (1990). Demisyllables as sets of features: Comments on Clement's paper. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I. Between the grammar and physics of speech* (pp. 377–381). Cambridge University Press.

Garrett, M. (1975). The analysis of sentence production. In G. Bower (Ed.), *The psychology of learning and motivation* (pp. 133–177). Academic Press. https://doi.org/10.1016/S0079-7421(08)60270-4

Gauvin, H., De Baene, W., Brass, M., & Hartsuiker, R. (2016). Conflict monitoring in speech processing: An fMRI study of error detection in speech production and perception. *NeuroImage*, *126*, 96–105. https://doi.org/10.1016/j.neuroimage.2015.11.037

Gauvin, H., & Hartsuiker, R. (2020). Towards a new model of verbal monitoring. *Journal of Cognition*, *3*(1), 1–37. https://doi.org/10.5334/joc.81

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. Arbib (Ed.), *Action to language via the mirror neuron system* (pp. 215–249). Cambridge University Press. https://doi.org/10.1017/CBO9780511541599.008

Goldstein, L., & Fowler, C. (2003). Articulatory phonology: A phonology for public language use. In N. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production* (pp. 159–208). De Gruyter Mouton. https://doi.org/10.1515/9783110895094.159

Goldstein, L., Nam, H., Saltzman, E., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. In G. Fant, H. Fujisaki, & J. Shen (Eds.), *Frontiers in phonetics and speech science: Festschrift for Wu Zongji* (pp. 239–249). Commercial Press.

Grandchamp, R., Rapin, L., Perrone-Bertolotti, M., Pichat, C., Haldin, C., Cousin, E., Lachaux, J.-P., Dohen, M., Perrier, P., Garnier, M., Baciu, M., & Lœvenbruck, H. (2019). The ConDialInt model: Condensation, dialogality, and intentionality dimensions of inner speech within a hierarchical predictive control framework. *Frontiers in Psychology*, *10*, 2019. https://doi.org/10.3389/fpsyg.2019.02019

Hansen, S., McMahon, K., & de Zubicaray, G. (2019). Neural mechanisms for monitoring and halting of spoken word production. *Journal of Cognitive Neuroscience*, *31*(12), 1946–1957. https://doi.org/10.1162/jocn_a_01462

Hartsuiker, R., & Kolk, H. (2001). Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive Psychology*, *42*(2), 113–157. https://doi.org/10.1006/cogp.2000.0744

Hickok, G. (2014a). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, *29*(1), 2–20. https://doi.org/10.1080/01690965.2013.834370

Hickok, G. (2014b). Towards an integrated psycholinguistic, neurolinguistic, sensorimotor framework for speech production. *Language, Cognition and Neuroscience*, *29*(1), 52–59. https://doi.org/10.1080/01690965.2013.852907

Huey, E. (1968). *The psychology and pedagogy of reading*. MIT Press. (Original work published 1908).

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology*, *2*, 255. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1-2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Kelso, J., Saltzman, E., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, *14*(1), 29–59. https://doi.org/10.1016/S0095-4470(19)30608-4

Laganaro, M. (2019). Phonetic encoding in utterance production: A review of open issues from 1989 to 2018. *Language, Cognition and Neuroscience*, *34*(9), 1193–1201. https://doi.org/10.1080/23273798.2019.1599128

Langland-Hassan, P., & Vicente, A. (2018). *Inner speech: New voices*. Oxford University Press. https://doi.org/10.1093/oso/9780198796640.001.0001

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Levelt, W. (1989). *Speaking: From intention to articulation*. MIT Press.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary? *Cognition*, *50*(1-3), 239–269. https://doi.org/10.1016/0010-0277(94)90030-2

Lukatela, G., Eaton, T., Lee, C., & Turvey, M. (2001). Does visual word identification involve a sub-phonemic level? *Cognition*, *78*(3), B41–B52. https://doi.org/10.1016/S0010-0277(00)00121-9

Lukatela, G., Eaton, T., Sabadini, L., & Turvey, M. (2004). Vowel duration affects visual word identification: Evidence that the mediating phonology is phonetically informed. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(1), 151–162. https://doi.org/10.1037/0096-1523.30.1.151

Lukatela, G., & Turvey, M. (1994a). Visual lexical access is initially phonological: 1. Evidence from associative priming by words, homophones, and pseudohomophones. *Journal of Experimental Psychology: General*, *123*(2), 107–128. https://doi.org/10.1037/0096-3445.123.2.107

Lukatela, G., & Turvey, M. (1994b). Visual lexical access is initially phonological: 2. Evidence from associative priming by homophones, and pseudohomophones. *Journal of Experimental Psychology: General*, *123*(4), 331–353. https://doi.org/10.1037/0096-3445.123.4.331

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(4), 499–511. https://doi.org/10.1017/S0140525X98001265

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*(2), 452–460. https://doi.org/10.3758/BRM.42.2.452

Manoiloff, L., Segui, J., & Hallé, P. (2013). L'effet de fréquence dans l'accès aux propriétés phonologiques des noms d'objets [Frequency effect in accessing picture names' phonological properties]. *L'Année Psychologique*, *113*(3), 335–348. https://doi.org/10.4074/S0003503313003023

Manoiloff, L., Segui, J., & Hallé, P. (2016). Subliminal repetition primes help detecting phonemes in a picture: Evidence for

a phonological level of the priming effects. *Quarterly Journal of Experimental Psychology*, *69*(1), 24–36. https://doi.org/10.1080/17470218.2015.1018836

Marslen-Wilson, W. (1990). Activation, competition, and frequency in lexical access. In G. Altman (Ed.), *Cognitive models of speech processing* (pp. 148–172). MIT Press.

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review*, *101*(4), 653–675. https://doi.org/10.1037/0033-295X.101.4.653

McNeill, D., & Lindig, K. (1973). The perceptual reality of phonemes, syllables, words, and sentences. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 419–430. https://doi.org/10.1016/S0022-5371(73)80020-9

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*(3), 298–305. https://doi.org/10.1016/S0022-5371(81)90450-3

Meringer, R., & Mayer, K. (1895). *Versprechen und Verlesen: eine psychologisch-linguistische Studie*. Goschensche Verlag. (Reissued with introductory essay by A. Cutler & D. Fay, 1978. Amsterdam: John Benjamins). https://doi.org/10.1075/cipl.2

Meyer, A., Roelofs, A., & Brehm, L. (2019). Thirty years of speaking: An introduction to the special issue. *Language, Cognition and Neuroscience*, *34*(9), 1073–1084. https://doi.org/10.1080/23273798.2019.1652763

Mitterer, H., & McQueen, J. (2009). Processing reduced word-forms in speech perception using probabilistic knowledge about speech production. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 244–263. https://doi.org/10.1037/a0012730

Montani, V., Chanoine, V., Grainger, J., & Ziegler, J. C. (2019). Frequency-tagged visual evoked responses track syllable effects in visual word recognition. *Cortex*, *121*, 60–77. https://doi.org/10.1016/j.cortex.2019.08.014

Morgan, J., & Wheeldon, L. (2003). Syllable monitoring in internally and externally generated English words. *Journal of Psycholinguistic Research*, *32*(3), 269–296. https://doi.org/10.1023/A:1023591518131

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, I. Chitoran, C. Coupé, & E. Marsico (Eds.), *Approaches to phonological complexity* (pp. 299–328). Walter de Gruyter. https://doi.org/10.1515/9783110223958.297

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. De Gruyter.

Nooteboom, S., & Quené, H. (2017). Self-monitoring for speech errors: Two-stage detection and repair with and without auditory feedback. *Journal of Memory and Language*, *95*, 19–35. https://doi.org/10.1016/j.jml.2017.01.007

Nooteboom, S., & Quené, H. (2019). Temporal aspects of self-monitoring for speech errors. *Journal of Memory and Language*, *105*, 43–59. https://doi.org/10.1016/j.jml.2018.11.002

Norris, D., & Cutler, A. (1988). The relative accessibility of phonemes and syllables. *Perception and Psychophysics*, *43*(6), 541–550. https://doi.org/10.3758/BF03207742

Nozari, N., Dell, G., & Schwartz, M. (2011). Is comprehension necessary for error detection? A conflict-based account of monitoring in speech production. *Cognitive Psychology*, *63*(1), 1–33. https://doi.org/10.1016/j.cogpsych.2011.05.001

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168. https://doi.org/10.1121/1.1909864

Oppenheim, G., & Dell, G. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*(1), 528–537. https://doi.org/10.1016/j.cognition.2007.02.006

Oppenheim, G., & Dell, G. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory and Cognition*, *38*(8), 1147–1160. https://doi.org/10.3758/MC.38.8.1147

Özdemir, R., Roelofs, A., & Levelt, W. (2007). Perceptual uniqueness point effects in monitoring internal speech. *Cognition*, *105*(2), 457–465. https://doi.org/10.1016/j.cognition.2006.10.006

Perfetti, C., Bell, L., & Delaney, S. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language*, *27*(1), 59–70. https://doi.org/10.1016/0749-596X(88)90048-4

Perrone-Bertolotti, M., Rapin, L., Lachaux, J.-P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, *261*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034

Pitt, M., & Samuel, A. (1995). Lexical and sublexical feedback in auditory word recognition. *Cognitive Psychology*, *29*(2), 149–188. https://doi.org/10.1006/cogp.1995.1014

Port, R. (1981). Linguistic timing factors in combination. *The Journal of the Acoustical Society of America*, *69*(1), 262–274. https://doi.org/10.1121/1.385347

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*(2), 97–132. https://doi.org/10.1016/S0010-0277(00)00090-1

Price, C. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062

R development core team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org

Reisberg, D., Smith, J., Baxter, D., & Sonenshine, M. (1989). "Enacted" auditory images are ambiguous; "pure" auditory images are not. *The Quarterly Journal of Experimental Psychology Section A*, *41*(3), 619–641. https://doi.org/10.1080/14640748908402385

Saltzman, E., & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Sato, M., Baciu, M., Lœvenbruck, H., Schwartz, J.-L., Cathiard, M., Segebarth, C., & Abry, C. (2004). Multistable representation of speech forms: A functional MRI study of verbal transformations. *Neuroimage*, *23*(3), 1143–1151. https://doi.org/10.1016/j.neuroimage.2004.07.055

Savin, H., & Baver, T. (1970). The non perceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, *9*(3), 295–302. https://doi.org/10.1016/S0022-5371(70)80064-0

Schiller, N., Jansma, B., Peters, J., & Levelt, W. (2006). Monitoring metrical stress in polysyllabic words. *Language and Cognitive*

Processes, *21*(1-3), 112–140. https://doi.org/10.1080/01690960400001861

Sebastián-Gallés, N., Marti, M., Cuetos, F., & Carreiras, M. (2000). *Lexesp: una base de datos informatizada del español*. Universidad de Barcelona.

Segui, J., Dupoux, E., & Mehler, J. (1990). The role of the syllable in speech segmentation, phoneme identification, and lexical access. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 263–280). MIT Press.

Segui, J., Frauenfelder, U., & Mehler, J. (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology*, *72*(4), 471–477. https://doi.org/10.1111/j.2044-8295.1981.tb01776.x

Shattuck-Hufnagel, S. (1979). Speech errors as evidence for a serial order mechanism in sentence production. In W. Cooper & E. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295–342). Erlhaum.

Shatzman, K., & McQueen, J. (2006). Segment duration as a cue to word boundaries in spoken-word recognition. *Perception and Psychophysics*, *68*(1), 1–16. https://doi.org/10.3758/BF03193651

Smith, J., Wilson, M., & Reisberg, D. (1995). The role of subvocalization in auditory imagery. *Neuropsychologia*, *33*(11), 1433–1454. https://doi.org/10.1016/0028-3932(95)00074-D

Snodgrass, J., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Snoeren, N. D., Segui, J., & Hallé, P. A. (2008). On the role of regular phonological variation in lexical access: Evidence from voice assimilation in French. *Cognition*, *108*(2), B512–B521. https://doi.org/10.1016/j.cognition.2008.02.008

Sokolov, A. (1972). *Inner speech and thought*. Springer-Verlag. https://doi.org/10.1007/978-1-4684-1914-6

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, *48*(2), 233–254. https://doi.org/10.1016/S0749-596X(02)00513-2

Spinelli, E., Welby, P., & Schaegis, A. (2007). Fine-grained access to targets and competitors in phonemically identical spoken sequences: The case of French elision. *Language and Cognitive Processes*, *22*(6), 828–859. https://doi.org/10.1080/01690960601076472

Stevens, K., & Blumstein, S. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368. https://doi.org/10.1121/1.382102

Streeter, L., & Nigro, G. (1979). The role of medial consonant transitions in word perception. *The Journal of the Acoustical Society of America*, *65*(6), 1533–1541. https://doi.org/10.1121/1.382917

Swinney, D. A., & Prather, P. (1980). Phonemic identification in a phoneme monitoring experiment: The variable role of uncertainty about vowel contexts. *Perception and Psychophysics*, *27*(2), 104–110. https://doi.org/10.3758/BF03204296

Tian, X., & Poeppel, D. (2012). Mental imagery of speech: Linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, *6*, 314. https://doi.org/10.3389/fnhum.2012.00314

Tian, X., Zarate, J., & Poeppel, D. (2016). Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, *77*, 1–12. https://doi.org/10.1016/j.cortex.2016.01.002

Vanderwolf, C. (1998). Brain behavior, and mind: What do we know and what can we know? *Neuroscience and Biobehavioral Reviews*, *22*(2), 125–142. https://doi.org/10.1016/S0149-7634(97)00009-2

Van Orden, G. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory and Cognition*, *15*(3), 181–198. https://doi.org/10.3758/BF03197716

Van Turennout, M., Hagoort, P., & Brown, C. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science*, *280*(5363), 572–574. https://doi.org/10.1126/science.280.5363.572

Vigliocco, G., & Hartsuiker, R. J. (2002). The interplay of meaning, sound, and syntax in sentence production. *Psychological Bulletin*, *128*(3), 442–472. https://doi.org/10.1037/0033-2909.128.3.442

Vroomen, J., & de Gelder, B. (1994). Speech segmentation in Dutch: No role for the syllable. In *Proceedings of the 1994 International Conference on Spoken Language Processing* (pp. 1135–1138).

Warren, R. (1961). Illusory changes of distinct speech upon repetition – the verbal transformation effect. *British Journal of Psychology*, *52*(3), 249–258. https://doi.org/10.1111/j.2044-8295.1961.tb00787.x

Whalen, D. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception and Psychophysics*, *35*(1), 49–64. https://doi.org/10.3758/BF03205924

Whalen, D. (1991). Subcategorical phonetic mismatches and lexical access. *Perception and Psychophysics*, *50*(4), 351–360. https://doi.org/10.3758/BF03212227

Wheeldon, L., & Levelt, W. (1995). Monitoring the time-course of phonological encoding. *Journal of Memory and Language*, *34*(3), 311–334. https://doi.org/10.1006/jmla.1995.1014

Wheeldon, L., & Morgan, J. (2002). Phoneme monitoring in internal and external speech. *Language and Cognitive Processes*, *17*(5), 503–535. https://doi.org/10.1080/01690960143000308

Ziegler, J., & Jacobs, A. (1995). Phonological information provides early sources of constraint in the processing of letter strings. *Journal of Memory and Language*, *34*(5), 567–593. https://doi.org/10.1006/jmla.1995.1026

Zwitserlood, P., Schriefers, H., Lahiri, A., & van Donselaar, W. (1993). The role of syllables in the perception of spoken Dutch. *Journal of Experimental Psychology: Learning Memory and Cognition*, *19*(2), 260–271. https://doi.org/10.1037/0278-7393.19.2.260