

**Multistable Syllables as Enacted Percepts:  
A Source of an Asymmetric Bias in the Verbal Transformation Effect**

Marc Sato\*, Jean-Luc Schwartz, Christian Abry, Marie-Agnès Cathiard and H el ene L evenbruck

*Institut de la Communication Parl ee, CNRS UMR 5009*

*Institut National Polytechnique de Grenoble, Universit e Stendhal*

Total: 12022 words

\* Corresponding author.

Institut de la Communication Parl ee – UMR CNRS 5009

Institut National Polytechnique de Grenoble - Universit e Stendhal

46, avenue F elix Viallet, 38031 Grenoble Cedex 01 - France.

Office: +33 (0)4 76 57 48 27

Fax: +33 (0)4 76 57 47 10

E-mail address: sato@icp.inpg.fr

ABSTRACT

Perceptual changes are experienced while repeating a speech form in a rapid and continuous way, an auditory illusion known as the Verbal Transformation Effect. While verbal transformations are mostly considered to reflect the perceptual organization and interpretation of speech, the present study was designed to test whether speech production constraints may participate in the emergence of verbal representations.

To this aim, we examined whether variations in the articulatory cohesion of repeated nonsense words, specifically temporal relationships between articulatory-acoustic events, could lead to perceptual asymmetries in verbal transformations. A first experiment displayed variations in timing relations between two consonantal gestures embedded in various nonsense syllables in a repetitive speech production task. In a second experiment, French subjects repeatedly uttered these syllables with an active search for verbal transformation. Syllable transformation frequencies followed the temporal clustering between consonantal gestures: The more synchronized the gestures, the more stable and attractive the syllable. In a third experiment involving a covert repetition mode, the pattern was maintained without external speech movements. However, when using a pure perceptual condition in a fourth experiment, the previously observed perceptual asymmetries of verbal transformations disappeared.

These experiments demonstrate the existence of an asymmetric bias in the Verbal Transformation Effect linked to articulatory control constraints. The persistence of this effect from an overt to a covert repetition procedure provides evidence that articulatory stability constraints originating from the action system may be involved in auditory imagery. The absence of the asymmetric bias during a purely auditory procedure discards perceptual mechanisms as

possible explanations of the observed asymmetries.

Keywords: Verbal Transformation Effect, Auditory Imagery, Speech Perception and Production, Articulatory Temporal Clustering.

## INTRODUCTION

In a study on visual imagery, Chambers and Reisberg (1985) defended the view that "mental images in any modality have no existence outside our understanding of them, making the image and its comprehension inseparable. In perception, there is a physical stimulus, existing independently of the perceiver, which needs interpretation. However, in imagery there is no free-standing icon waiting to be interpreted, and no interpretation is needed to learn what the image depicts" (cited in Reisberg, Smith, Baxter and Sonenshine, 1989, p. 620). To verify this hypothesis, the authors carried out a set of experiments in which subjects were asked to imagine standard ambiguous figures, like the Necker cube. They showed that subjects were uniformly unable to mentally discover other shapes than the one provided by the experimenter. Following the experiment, subjects were however able to draw the figure and to discover interpretations different from the first one. Chambers and Reisberg concluded that visual mental images are inherently unambiguous.

Considering that auditory imagery could differ from visual imagery in the stimuli and in the perceptual properties of the two modalities, Reisberg et al. (1989) attempted to examine ambiguous images in the auditory domain. To this aim, they exploited the Verbal Transformation Effect (Warren and Gregory, 1958). This word game (Treiman, 1983), which bears an analogy with the depth perceptual rivalry in the Necker cube, relies on the fact that certain words, if repeated over and over, yield a soundstream compatible with more than one segmentation. For example, rapid repetitions of the word "life" may perceptually switch into sequences of the word "fly". Reisberg and colleagues (1989) used this verbal transformation paradigm to test the unambiguity assumption on mental images in the auditory domain by examining whether

imagined repetitions could produce verbal transformations, just as heard repetitions do. In their experiments, the authors requested the subjects to imagine the repetition of a word and to report any transformation of the auditory image. In order to test their assumption more thoroughly, they asked other subjects to detect possible transformations during the overt repetition of the same word produced either by the experimenter or by themselves.

However, the authors noted that in the imagined condition, subjects might supplement the auditory image with a subvocalized "enactment". According to the authors, enactment could provide real physical cues that might help auditory transformations, while "pure" auditory imagery would not. To control for this potential shortcoming in the paradigm, the authors tested various conditions differing on the degree of enactment (whispering, silent mouthing, imaging with no mouthing) or even eliminating enactment (with a concurrent articulatory task, or by chewing candy, or by clamping the articulators). The experimental data showed that transformations were largely eliminated when subarticulation was blocked. Eliminating enactment, either by concurrent articulation or by clamping the articulators, prevented subjects from detecting a transformation. Therefore, transformations in auditory imagery seem to require subvocalization. Moreover, the authors found that the transformation probability gradually decreases from a condition of complete externalization to one of complete internalization, through a condition of partial externalization (whispering, mouthing).

Reisberg et al. (1989) concluded that subvocalized enactment enables refreshment and, thus, elaboration of verbal auditory images, whereas "pure" unenacted auditory images remain unambiguous, just as visual images are. They thereby provided the first demonstration that speech production constraints, specifically speech enactment, may intervene in the emergence of

verbal transformations.

Verbal transformations: A window into speech representations

In past research, verbal transformations have been mainly studied as a pure perceptual effect. The classical paradigm consists in presenting subjects with an auditory speech stimulus looped on a tape (e.g., Warren, 1961, 1982; Warren and Meyers, 1987; Kaminska, Pool and Mayer, 2000; Pitt and Shoaf, 2001, 2002; Shoaf and Pitt, 2002). It has been shown that the number of transformations heard by listeners depends on the stimulus length, the interstimulus interval duration and the listening duration (Warren, 1961). Previous studies have reported that perceptual changes, when compared to the auditory input, could range from small phonetic deviations to strong semantic distortions, including substitution of a phoneme by a phonetically close one (Warren, 1961; Warren and Meyers, 1987), auditory streaming/perceptual grouping (Pitt and Shoaf, 2001, 2002), lexical and semantic transformations (Warren, 1961; Kaminska, Pool and Mayer, 2000, Shoaf and Pitt, 2002). Lexical and sublexical levels of representation have been suggested as the loci of such effects. Accordingly verbal transformations should vary as a function of distinct factors related to the repeated stimulus: its neighborhood density (i.e., the number of lexical entries that are phonologically similar to the repeated stimulus), its frequency in the language and whether it is a word or not (Natsoulas, 1965; Yin and MacKay, 1992; MacKay et al., 1993; Shoaf and Pitt, 2002).

Although the processes implied in verbal transformations have challenged unified theoretical explanation for more than four decades, these transformations are mostly considered to reflect the operation of processes devoted to the perceptual organization and interpretation of speech (Warren, 1982). Two functions seem to be involved in the reinterpretation of the signal

when it does no longer make sense: satiation and criterion shift (Warren and Meyer, 1987; MacKay, Wulf, Yin and Abrams, 1993; Kaminska, Pool and Mayer, 2000). The repeated listening of a stimulus causes its memory representation to satiate. Simultaneously, the criteria used to categorize it abruptly shift and a new representation is then built. These processes would repeat themselves throughout the stimulus presentation.

### Articulatory Synchrony Hypothesis

Viewed as temporary fluctuations of the on-line linguistic information processing that operate during veridical perception, auditory illusions provide a useful framework to enhance our understanding of the language system by revealing otherwise hidden mechanisms (Warren, 1982). From this point, the Verbal Transformation Effect appears to be well suited to explore the organization of lexical and sublexical representations by examining variations in the perceptual stability of reported transformations. Furthermore, the fact that the effect occurs not only during a purely auditory procedure but also during an overt or a covert repetition condition and that speech production constraints, specifically speech enactment, could also intervene in the transformation process (Reisberg et al., 1989; Smith et al., 1995) makes the Verbal Transformation Effect a nice pivot point to examine whether speech production and perception constraints act on verbal auditory images. In this framework, the aim of the present study was to further test how *specific* articulatory constraints may contribute to verbal transformations and to examine the functional equivalence, in terms of transformation mechanisms, between the classic perception procedure and the production variant procedure.

In this regard, an important issue neither considered by Reisberg et al. (1989) nor in other verbal transformation studies, concerns the existence of possible verbal transformation

asymmetries. For instance, the reverse transformation from "fly" to "life" seems less likely than from "life" to "fly". This could be related to differences in the articulatory cohesion of speech stimuli (Browman and Goldstein, 1989): The two consonantal gestures in the syllable onset of "fly" are temporally very compact and the speaker can produce the three gestures of /flai/ almost in synchrony. In the "life" sequence, the synchronization of [l] (syllable onset) and [f] (syllable coda) is of course impossible. Hence, the temporal clustering in /flai/, could explain the facilitation for the transformation from "life" to "fly".

More generally, we suggest that transformations from less to more temporally clustered stimuli should be more frequent, since they are associated with more compact and tightly synchronized sequences of articulatory gestures. From this view and since the pioneering work by Stetson (Stetson 1951), several studies have reported evidence for variations in articulatory phase relationship during speech production (e.g., Tuller and Kelso, 1990, 1991; Gleason, Tuller and Kelso, 1996; de Jong, 2001; de Jong, Nagao and Lim, 2002). By using a repetitive speech production task, Stetson (1951) first observed that fast repetition rates could induce specific syllabic parsing (i.e., the resyllabification of codas in VC vowel-consonant sequences into onsets in CV consonant-vowel sequences). Tuller and Kelso (1990, 1991) further replicated this finding by introducing the concept of relative phasing of articulatory events. They showed that, at increasing speaking rate, while the /pi/ sequence remained stable all along the task, they observed a switch from /ip/ to /pi/ at a critical rate (see also Gleason, Tuller and Kelso, 1996; de Jong, 2001; de Jong, Nagao and Lim, 2002). The greater stability of the CV phonetic structure as compared to the VC one is in accordance with studies of intrasyllabic tendencies during the babbling and single-word periods in early vocal acquisition (MacNeilage, 1998; MacNeilage and

Davis, 2000) and is also supported by the cross-linguistic typology literature showing a clear predominance of CV syllables (e.g., Jakobson, 1966; Maddieson, 1984). Hence, the finding of variations in phasing patterns has provided a useful framework to rationalize a number of typological facts.

Since the Verbal Transformation Effect is well suited for assessing perceptual stability and phonological consciousness, the present study was first designed to test whether variations in the articulatory cohesion of repeated nonsense syllables could lead to perceptual asymmetries in verbal transformations. To this aim, the first experiment was designed to select a convenient phonetic material and display variations in the temporal clustering of articulatory-acoustic events within this material. The second experiment, using an overt repetition mode, was designed to test whether these variations could lead to perceptual asymmetries in the reported verbal transformations. The aim of Experiment 3 was to further explore such possible perceptual asymmetries using a covert repetition mode and then to examine whether synchrony constraints originating from the action system may participate in the elaboration of verbal auditory images. Finally, the aim of Experiment 4, involving a pure perceptual condition, was to disentangle the role of perception mechanisms and production constraints in the pattern of results provided by the previous experiments.

## EXPERIMENT 1

In order to explore possible perceptual asymmetries in the Verbal Transformation Effect, a convenient set of speech stimuli was first selected, likely to display various degrees of articulatory cohesion. The aim of the first experiment was to study variations in timing relations of articulatory-acoustic events within this material during a repetitive speech production task at various rates. This is in line with a large body of literature on speech timing in which reorganisations in timing at high rates provide an implicit attractor to the speech production system, and shed light on the respective cohesion of various competing sequences.

### Method

#### *Phonetic Material*

Six monosyllabic nonsense words were selected. Each sequence – i.e., /psə/, /səp/, /əps/, /spə/, /pəs/ and /əsp/ - consisted of a combination of the bilabial [p] and coronal [s] consonants and the neutral vowel [ə]. The neutral vowel [ə] was selected since it provides minimal constraints on the vocal tract shape, and hence lets the articulators for /p/ and /s/ free for achieving their consonantal task.

None of these syllables occurs as a word in the French lexicon, allowing to minimize lexical interferences in the verbal transformation task (Shoaf and Pitt, 2002). However, once discarded the neutral vowel, the phonological types of the speech sequences (i.e., /psV/, /sVp/, /Vps/, /spV/, /pVs/ and /Vsp/ - V being any oral French vowel) are all phonotactically attested in French. From this point, it is also important to note that [ə] is generally realized as a mid-open front rounded vowel, making its articulatory realization close to [œ], which can occur both in

close and open syllables in French.

Articulatory-acoustic Events

The consonants and the vowel were characterized by onset acoustic events with clear articulatory interpretation. Schematic unfolding of onset events for the six sequences are shown in Figure 1. For /səp/, the frication onset for [s] is followed by the voicing onset for [ə] and the [p] release after the bilabial closure. For /psə/, two of these events more or less cohere into a single event, defined by the onset of friction at the beginning of the logatome, and corresponding to tightly synchronous gestures for [p] (lip opening) and [s] (tongue tip placing). For /əps/, these two events still cohere, but this happens long after the voicing onset for [ə], and after the lip closure for [p]. The three individual onset events for [s], [p] and [ə] are all separated in isolated /spə/, /pəs/ and /əsp/ sequences. For /spə/ or /pəs/ cycles, the two onset events for [s] and [p] remain separated by the lip closure period, hence they never cohere. For /əsp/, the frication onset and the bilabial release events happen only after the voicing onset for [ə].

---



---

Insert Figure 1 about here

---



---

In order to test the stability and cohesion of sequences during a repetitive speech production task, we focused on the relative timing of the onset events for the two consonantal gestures. As described above, these events should remain always separated in cycles for /spə/ and /pəs/. For /psə/ and /səp/ sequences, an important question is to know if the onset events

might cohere in cycles for /səp/, the release of the final [p] becoming synchronous with the tongue-driven onset of the initial [s] in the next /səp/ utterance. This would result in the resyllabification from /səp/ to /psə/.

### Apparatus and Procedure

Examination of /səp/ cycles showed that [p] becomes implosive almost at once. This means that the [p] burst disappears, hence it becomes impossible to track the /ps/ coherence on the acoustic signal, since the [p] onset event is no longer noticeable. To escape from this difficulty and to test the reality of coordinative patterns of onset events, we performed *audio-visual* recordings of /psə/, /səp/, /pəs/ and /spə/ sequences. The idea was that we suspected [p] to be actually released on the lips, even with no acoustic noise. We used the set-up designed at the Speech Communication Institute for such phonetic analyses, in which the speaker's lips are coloured with blue make-up, to allow precise video analyses using a chroma-key process (Lallouache, 1990).

The four sequences were individually recorded by a trained phonetician, native speaker of French (J.-L.S.). Two rate manipulations were examined. In an increased rate condition, the speaker progressively speeded up the rhythm, from a low rhythm (around 1 cycle per second) to a high one (around 6 cycles per second). In a fixed rate condition, the speaker uttered the cycles at two stable rhythms, respectively low (around 1 cycle per second) or moderate (around 2 cycles per second). Though previous studies on articulatory phasings always used a speeding-up speech production paradigm (Tuller and Kelso, 1990, 1991; Gleason, Tuller and Kelso, 1996; de Jong,

2001; de Jong, Nagao and Lim, 2002), the fixed rate condition was designed to test if variations in timing relationships between articulatory-acoustic events would occur at a sufficient rate, but without any acceleration.

All the stimuli were then analysed in the following way. The initiation of the high-frequency noise characteristic of the [s] onset event was detected on the stimuli spectrogram (using the Praat software, Institute of Phonetic Sciences, University of Amsterdam). The [p] onset event was localised by analysing the variations of the lip area and by detecting the first video frame with a non-zero area after the closure period (the system may detect lip areas as small as 1 mm<sup>2</sup>, with a temporal precision of 20 ms, Abry et al., 2004). Then, we defined the /ps/ coherence index as the time separating the onset events for [p] and [s], divided by the cycle duration defined as the time between two consecutive [p] events.

### Results

On Figure 2.1 are plotted the variations of the index as a function of the cycle duration (in ms) for the /psə/ and /səp/ speech sequences in the increased rate condition. Whichever the speech rate, the index was stable around zero for /psə/. While the index for /səp/ was around .50 at low speed (around 1 cycle per second), it abruptly decreased towards zero from 2 cycles per second. We observed a similar pattern for the two speech sequences in the fixed rate condition (see Figure 2.2). Whichever the speech rate (i.e., low or moderate), the index was stable around zero for /psə/ (mean index = 0.02, *SD* = 0.01 in the low speech rate condition; mean index = 0.03, *SD* = 0.01 in the moderate speech rate condition). At a low speech rate, the index for /səp/ was around .40-.50 (mean index = 0.43, *SD* = 0.02), while it was around .10-.20 during the

moderate speech rate condition (mean index = 0.15,  $SD = 0.06$ ). A two-way analysis of variance (ANOVA), with the type of speech sequence (*stimulus condition*) and the speech rate (*speech rate condition*; i.e., low or moderate) as independent variables and the coherence index as the dependant variable, yielded a significant effect of the stimulus condition [ $F(1,32) = 620.79$ ,  $MSe = 0.65$ ,  $p < .001$ ], a significant effect of the speech rate condition [ $F(1,32) = 159.76$ ,  $MSe = 0.17$ ,  $p < .001$ ], and a reliable interaction between the two conditions [ $F(1,32) = 186.32$ ,  $MSe = 0.19$ ,  $p < .001$ ].

---



---

Insert Figures 2.1 & 2.2 about here

---



---

The pattern for /pəs/ and /spə/ is different. In the increased rate condition (see Figure 3.1) the indexes of these sequences got gradually closer from a speech rate of around 2 cycles per second. In the fixed rate condition (see Figure 3.2), for the low speech rate condition, the index was around .70-.80 for /pəs/ (mean index = 0.76,  $SD = 0.01$ ) and around .20-.30 for /spə/ (mean index = 0.25,  $SD = 0.02$ ). When cycle duration decreased and rate increased (around 2 cycles per second), we observed a nearly identical index around .45-.60 for both speech sequences (mean index = 0.58,  $SD = 0.02$  for /pəs/; mean index = 0.47,  $SD = 0.01$  for /spə/). Hence, whichever the repetition rate, [s] and [p] always remain separated in cycles for /spə/ and /pəs/. Furthermore, for a rhythm of around 2 cycles per second, /pəs/ and /spə/ sequences are barely distinguishable in terms of acoustic events. A two-way ANOVA yielded a significant effect of the stimulus condition [ $F(1,32) = 3046.85$ ,  $MSe = 0.85$ ,  $p < .001$ ], a significant effect of the speech rate condition [ $F(1,32) = 14.97$ ,  $MSe = 0.00$ ,  $p < .001$ ], and a reliable interaction between the two

conditions [ $F(1,32) = 1328.84$ ,  $MSe = 0.37$ ,  $p < .001$ ].

---



---

Insert Figures 3.1 and 3.2 about here

---



---

### Discussion

In summary, the data for /psə/ and /səp/ sequences fit well with the claim that more temporally clustered stimuli should be more stable and play the role of attractors during a repetitive speech task. Analyses of timing relations between articulatory-acoustic events during a repetitive speech production task showed that, at a low rhythm, the onset events appear to be more synchronous for /psə/ than for /səp/. In fact, for the /psə/ sequence, consonants in the syllable onset are temporally clustered and hence may be thought of as a tightly synchronized unit (for further evidence on the clustering of CV structures, see MacKay, 1974; de Jong, 2003). At a rhythm above 2 cycles per second, /psə/ should play the role of an attractor for cycling /səp/, the release of the final [p] becoming synchronous with the tongue-driven onset of the initial [s] in the next utterance. Whichever the repetition rate, [s] and [p] onset events never cohere in cycles for /spə/ and /pəs/, since they are separated by lip closure. However, /pəs/ and /spə/ sequences appear to be barely distinguishable in terms of articulatory-acoustic events at a rhythm above 2 cycles per second. Finally, variations in timing relationships between articulatory-acoustic events also occurred without any rate acceleration. This suggests that a sufficient speaking rate is more crucial than a speeding-up paradigm in the study of articulatory phasing *per se*.

EXPERIMENT 2

The purpose of Experiment 2 was to test the existence of preferential transformations by contrasting more or less “temporally clustered” syllable stimuli, during an overt production variant of the verbal transformation paradigm. As shown in Reisberg and colleagues’ study (1989), the efficiency of verbal transformations depends on the degree of enactment. Therefore, we first adopted the most simple and effective condition for eliciting transformations: i.e., overt repetition.

We assumed that the repetitive production of the speech stimuli might result in shifted sequences, belonging to two groups (i.e., in Group 1, the /psə/, /səp/ and /əps/ sequences and in Group 2, the /spə/, /pəs/ and /əsp/ sequences). In other words, we predicted that participants, when presented with a given sequence, would not produce all the possible permutations but would be naturally brought to “mentally read” the result of their repetition according to a “shifting” parsing within each group. For example, repetition of the /psə/ sequence would lead to a “shifting” segmentation where a perceptual boundary may be placed after [ə], [p] or [s].

Our claim is that more temporally clustered stimuli should be more stable and play the role of attractors in Verbal Transformations. The speech production data in Experiment 1 indicate that in Group 1, the /səp/ sequence should be transformed into /psə/, while in Group 2, /spə/ and /pəs/ should be equally stable. Furthermore, the assumption concerning VCC sequences in both groups is that they should be least stable, since the consonantal gesture(s) in the coda intervenes long after the vowel initiation. This results in the pattern of predictions in

Table 1, in which it is assumed that the lower the articulatory-acoustic coherence (or the less temporally clustered the sequence), the less stable the sequence, and the more likely its transformation into a more coherent one.

---

---

Insert Table 1 about here

---

---

### Method

#### Participants

Fifty-six students from Grenoble University participated in this experiment. All were native French speakers without hearing or speaking disorders, and were naive as to the purpose of the experiment.

#### Apparatus

For follow-up analyses, the experiment was entirely recorded onto a portable audio-recorder. The recording was then digitized as individual sound files to the hard disk of a PC computer at a sampling rate of 22.05 kHz with 16-bit quantization.

#### Procedure

The participants were tested individually. The experiment began with a lengthy briefing during which they were introduced with the verbal transformation task. The participants listened to the experimenter repeating the word "life" at a rate of two repetitions per second and were asked to listen carefully for any changes in the repeating utterance. The experimenter then asked the participants if they had perceived another sequence and, if not, explained the possibility of hearing the word "fly". This "bootstrap" example, presented in English rather than French, aimed at displaying the Verbal Transformation Effect on a material that all subjects understood, while

letting them experience the phenomenon later on their own production and in their own language. Then, they were told that they would repeat a given sequence aloud into the microphone placed in front of them, at a rate of about 2 cycles per second, with no gap between repetitions. If they heard a transformation, they were to stop and report it. If they did not hear any transformation, they were to say nothing; the experimenter would stop them after thirty seconds. Finally, it was indicated that possible changes could be subtle as well as very noticeable and could correspond to a word as well as a nonsense utterance. Furthermore, the participants were assured that there were no correct or incorrect responses.

In the test session, the six sequences – i.e., /psə/, /səp/, /əps/, /spə/, /pəs/ and /əsp/ - were orally presented by the experimenter in one of six counterbalanced orders (based on the sequence alternation from one group to the other, excluding the successive presentation of two sequences with a similar onset). If the subject made pronunciation mistakes, paused (thereby breaking rhythm), slowed down or stopped before the 30s time period without reporting any change, the experimenter asked him/her to resume the ongoing activity. Lengthy breaks were offered between trials.

### Results

The transformation frequencies observed for the six sequences and averaged over subjects are shown in Table 2a. The stimulus sequences are given in rows and the observed transformations are given in columns. Diagonal entries are the percentage of stable utterances while the off-diagonal entries correspond to transformed sequences.

---



---

Insert Table 2 about here

---



---

The observed transformations from one group to the other were extremely rare (on average 3% for the two groups). Column "misc." represents the percentage of unpredicted transformations. For 4% of the overall responses, they involved lexical transformations (such as /sø/ ("these"), /pø/ ("few") or /saspø/ ("it may be")), and, for 6% of the responses, they corresponded to a nonsense word with a larger phonological structure than expected (such as /psəp/ for /səp/). The total number of such transformations, although not trivial, remains however small (on average 11%) with respect to the fact that participants were not informed about expected transformations. Overall, most responses occurred within each group (on average, 86% of the responses), which confirms the shifting parsing hypothesis.

*Stabilities and preferential transformations*

Global statistical analyses yielded no significant effect of the six counterbalanced stimulus orders ( $\chi^2(5) = 4.39$ ,  $p$  value exceeded the .05 level, unless otherwise stated) and a significant global stimulus effect ( $\chi^2(5) = 32.10$ ,  $p < .0005$ ).

According to the observed shifting parsing process, further statistical comparisons were carried-out on each group separately. We tested discrepancies between sequences on two distinct measures. The stability index was calculated by summing the number of times a given sequence was not transformed. The attractivity index, evaluating the sequence capacity to attract, or "capture", the other sequences during the repetition process, was calculated by summing the number of times a given sequence was selected as a transformation within a group, weighted by the number of times it could have been selected as a transformation.

Within Group 1 (see Table 2b), the global comparison of the observed stability per

sequence yielded a significant effect ( $\chi^2(2) = 24.76, p < .0001$ ). Analyses of stability across sequences, with a Bonferroni correction (applied in all the following individual comparisons), showed reliable discrepancies between /psə/ and /səp/ ( $\chi^2(1) = 22.39$ ) and between /səp/ and /əps/ ( $\chi^2(1) = 12.93$ ). The global comparison within Group 1 of the observed attractivity per sequence was reliable ( $\chi^2(2) = 64.12, p < .0001$ ) with significant discrepancies between /psə/ and /səp/ ( $\chi^2(1) = 18.07$ ), between /psə/ and /əps/ ( $\chi^2(1) = 61.38$ ) and between /səp/ and /əps/ ( $\chi^2(1) = 14.21$ ). In summary, within Group 1, /psə/ and /əps/ showed a stronger stability than /səp/, whereas /psə/ was the most attractive sequence and /səp/ showed a stronger attractivity than /əps/. Within Group 2 (see Table 2c), the global comparison of the observed stability per sequence yielded no significant effect ( $\chi^2(2) = 3.51$ ). However, the global comparison of the observed attractivity per sequence was reliable ( $\chi^2(2) = 59.57, p < .0001$ ) with significant differences between /pəs/ and /əsp/ ( $\chi^2(1) = 59.49$ ) and between /spə/ and /əsp/ ( $\chi^2(1) = 39.78$ ). In summary, within Group 2, we observed no reliable discrepancies between sequences in terms of stability and a stronger attractivity for /pəs/ and /spə/ than for /əsp/.

#### Test of a glottal onset effect

The /əps/ and /əsp/ syllables with an empty onset, although predicted as very unstable, appeared rather stable in this experiment (although not at all attractive). This could be explained by the presence of a glottal stop often produced at the syllable onset by the participants. This glottal stop can be considered as an additional consonant in the syllabic structure, transforming

*/əps/* and */əsp/* VCC syllables into */ʔəps/* and */ʔəsp/* CVCC syllables, respectively. This glottal onset might prevent the fast and connected repetition of items and hence block articulatory synchronization (de Jong, 2001). Consistent with this hypothesis was the longer mean duration rate observed for the two sequences as compared to the others (we also observed a longer mean duration rate for */əsp/* as compared to */əps/*). A post-hoc phonetic analysis was conducted by two trained phoneticians, consisting in determining the presence or the absence of a glottal stop in the final portion of each of the */əps/* and */əsp/* recordings without indication about the observed stability/instability of the sequence. This analysis confirmed that 71% and 63% of the participants produced a glottal stop at the end of the repetition process for the */əps/* and */əps/* sequences, respectively. Further comparisons between transformed and untransformed sequences showed that the glottal onset was produced for 86% of the untransformed */əps/* sequences and for 45% of the transformed */əps/* sequences, and for 82% of the untransformed */əsp/* sequences and for 40% of the transformed */əsp/*. Reanalyses of results excluding participants who produced a glottal onset showed that the */əps/* sequence remained untransformed for 36% of the participants while it was transformed towards */psə/* or any unexpected transformation for 50% and 14% of the participants, respectively. In the same way, the */əsp/* sequence remained untransformed for 27% of the participants while it was transformed towards */pəs/*, */spə/*, */psə/* or unexpected transformations for 41%, 9%, 9% and 14% of the participants, respectively. Taken

together, these results thus confirm that the observed stabilities for /əps/ and /əsp/ were largely due to a glottal onset effect and explain why their respective stability and attractivity degrees differ so much (i.e., high stability vs. low attractivity).

### Discussion

Altogether, the results fit reasonably well with the expectations summarized in Table 1. Firstly, the shifting parsing seems to be the rule. Secondly, the hierarchy of attractivities within each group mirrors the proposed hierarchy of articulatory cohesion (i.e., "/psə/ > /səp/ > /əps/" in Group 1 and "/pəs/ = /spə/ > /əsp/" in Group 2). Stability patterns are less in agreement with our predictions, but the glottal onset effect is the major responsible factor for this, and, once taken into account, both the stability and attractivity patterns correspond with the predictions.

Considering the unexpected transformations, a number of previously underlined contents of transformations were observed in our experiments, including substitution of a phoneme by a phonetically close one (e.g., /səb/ for /səp/, or /əbs/ for /əps/ - Warren, 1961; Warren and Meyers, 1987), auditory streaming (e.g., /əp/ for /əps/ - Pitt and Shoaf, 2001, 2002) and lexical transformations (e.g., /sø/ ("these"), /pø/ ("few") or /saspø/ ("it may be") for /spə/, or /pus/ ("thumb") for /pəs/ - Warren, 1961; Kaminska, Pool and Mayer, 2000, Shoaf and Pitt, 2002). However, in the present experiment, the majority of the observed transformations cannot be related to such lexical or phonological transformation processes. By using a production variant of the classical verbal transformation paradigm, Reisberg and colleagues (1989) first demonstrated that speech production constraints, specifically speech enactment, could also intervene in the

transformation process. The observed asymmetries in the reported transformations reinforced this position by showing that the perceptual stability and attractivity of an uttered sequence might also depend on *specific* articulatory constraints - i.e., the temporal clustering between intrasyllabic articulatory gestures.

### EXPERIMENT 3

According to Reisberg and colleagues' results (1989), the decrease in enactment from overt to covert speech should result in a decrease in the number of transformations, the interpretation being that elaboration of verbal auditory images depends on the subvocalized enactment degree. The question is however to know if the specific articulatory constraints related to variations in temporal clustering of /p/ and /s/ are preserved in covert speech, and produce verbal transformations asymmetries as in the overt mode. The purpose of the following experiment was to further examine this hypothesis by testing the persistence of verbal transformations asymmetries in a covert repetition mode.

#### Method

##### *Phonetic materials*

The stimuli used in this experiment were the same as in Experiment 2. The assumptions about the "shifting parsing" and the articulatory cohesion hierarchy were therefore the same.

##### *Subjects and task*

Twenty-nine new participants were recruited from Grenoble University. All were native speakers of French without hearing or speaking disorders, and were naive as to the purpose of the experiment.

##### *Procedure*

The participants were tested individually. As in Experiment 2, they were firstly introduced with the verbal transformation task. Then, they were told that they would mentally repeat a given sequence, keeping their mouth closed, at a rate of about 2 cycles per second, with no gap between repetitions, and were asked to "mentally listen" for any changes in the repeating

utterance. If they found a transformation, they were to stop and report it. If they did not hear any transformations, they were to say nothing. It was indicated that possible changes could be subtle as well as very noticeable and could correspond to a word as well as a nonsense utterance. As previously, the participants were assured that there were no correct or incorrect responses.

In the test session, the six sequences – i.e., /psə/, /səp/, /əps/, /spə/, /pəs/ and /əsp/ - were orally presented by the experimenter in one of six counterbalanced orders. During the covert repetition, some participants happened to move their lips without phonation; in this case, the experimenter asked them to keep their mouth closed and to start again without moving the lips. Lengthy breaks were offered between trials.

### Results

---



---

Insert Table 3 about here

---



---

From the results (see Table 3a, same presentation as Table 2a), the percentage of the observed transformations from one group to the other were on average 11% while the percentage of unpredicted transformations was on average 13%. Most responses occurred within each group (on average, 76% of the responses), according to a shifting parsing process. This percentage was however 10% lower than in Experiment 2.

#### *Stabilities and preferential transformations*

A global statistical analysis of the results displayed no significant effect of stimulus order ( $\chi^2(5) = 1.74$ ) and a significant global stimulus effect ( $\chi^2(5) = 12.98, p < .05$ ). Within each group, the statistical comparisons of stability and attractivity patterns showed the following results. Within Group 1, the global comparison of the observed stability per sequence was not

significant ( $\chi^2(2) = 5.90$ ). The global comparison of the observed attractivity per sequence was reliable ( $\chi^2(2) = 43.39, p < .0001$ ) with significant differences between /psə/ and /səp/ ( $\chi^2(1) = 17.38$ ) and between /psə/ and /əps/ ( $\chi^2(1) = 25.83$ ). Within Group 2, the global comparison of the observed stability per sequence was significant ( $\chi^2(2) = 6.99, p < .05$ ) with /pəs/ reliably differing from /əsp/ ( $\chi^2(1) = 6.90$ ). The global comparison of the observed attractivity per sequence was reliable ( $\chi^2(2) = 10.81, p < .005$ ) with significant differences between /pəs/ and /əsp/ ( $\chi^2(1) = 6.55$ ) and between /spə/ and /əsp/ ( $\chi^2(1) = 8.10$ ). In summary, we observed within Group 1 no reliable discrepancies between sequences in terms of stability and a stronger attractivity for /psə/ than for /səp/ and /əps/. Within Group 2, we observed a stronger stability for /pəs/ than for /əps/ and a stronger attractivity for /pəs/ and /spə/ than for /əsp/.

### Discussion

In terms of the attractivity degree per sequence, the covert repetition mode explored in Experiment 3 produced patterns of verbal transformations asymmetries similar to those of the overt mode in Experiment 2 ( $\chi^2(5) = 5.75, N.S.$ ). Indeed, there is a convergence between the results of the two experiments showing a nearly similar hierarchy "/psə/ > /səp/ ≥ /əps/" within Group 1 and the same hierarchy "/pəs/ = /spə/ > /əsp/" within Group 2. Furthermore, these attractivity patterns mostly correspond with our predictions.

However, the patterns of stabilities significantly differ between the two experiments

( $\chi^2(5) = 13.20, p < .05$ ). These differences firstly come from the higher stability of the /əps/ and /əsp/ sequences in Experiment 2 (where they displayed an unexpectedly high stability) compared with Experiment 3 (where they presented a decreased stability, more in line with our predictions). As described previously, we explain the unpredicted stability of the /əps/ and /əsp/ sequences in Experiment 2 by the presence of a glottal stop often produced at the syllable onset, hence preventing fast and connected repetitions of items and then blocking articulatory synchronization (de Jong, 2001). Although the stability discrepancies of these sequences observed between the two experiments could be due to differences in subjects, we cannot exclude the possibility that the glottal effect might have decreased in Experiment 3, hence depending on the degree of external articulation. Considering the latter hypothesis, the fact that, under covert conditions, some articulatory control constraints - i.e., the temporal clustering between intrasyllabic articulatory gestures – would remain active in the building up of auditory images while the glottal onset effect would disappear might constitute an interesting phenomenon to further examine in the study of the functional equivalence between overt and covert speech. This would require additional tests which are under the scope of the present work.

Another source of stability discrepancies between the two experiments is the lower number of transformations in the covert repetition condition. Indeed, if the /əps/ and /əsp/ sequences are discarded because of possible discrepancies between subjects or glottal onset effect size, a higher stability of the sequences is observed in the covert repetition condition compared with the overt one. This is quite similar to what was observed by Reisberg et al. (1989): When contrasting all the displayed transformations, the authors found an average 38%

decrease (according to the transformation results of the monosyllabic word "stress" to "dress", see p. 635), while we found an average 26% decrease. Interestingly, the number of shifting parsing violations is quite larger in the covert repetition condition. Particularly /psə/ plays the role of an attractor for an important proportion of sequences in Group 2. This suggests that the segmental order of articulatory events of the repeated utterances is more difficult to maintain during a covert repetition, possibly due to fewer auditory and proprioceptive inputs in the control of the uttered sequence (Murray, 1965).

In summary, when contrasting the results of the two experiments, we observe a varying stability for the /əps/ and /əsp/ sequences between experiments and a stronger stability of the other sequences in the covert mode. It is therefore remarkable that, in spite of these differences, the attractivity pattern, which mostly corresponds with our articulatory cohesion predictions, is maintained from the overt to the covert repetition condition. This suggests that the temporal clustering of articulatory-acoustic events can also take place internally, without any external stimulus neither articulatory nor auditory. This result appears to be consistent with previous behavioural studies showing in some degree a functional equivalence between overt and covert speech (e.g., Landauer, 1962; MacKay, 1982; Postma and Noordanus, 1996) and, in a more general way, with the burgeoning domain of 'motor cognition' (providing strong empirical evidences for a functional coupling between simulated action and executed one - for a review, see Jeannerod, 1994). In line with these studies, the persistence of the asymmetric bias from an overt to a covert repetition procedure suggests that constraints from the speech production system seem able to penetrate verbal imagery and participate in the mental analysis and

interpretation of phonological forms during the emergence of verbal transformations.

#### EXPERIMENT 4

Speech is a matter of gestures and sounds, resulting in a set of more or less clustered events, which are both, to a certain extent, audible and articulatorily interpretable. Considering that verbal transformations may involve both perceptual mechanisms (e.g. auditory streaming) and motor constraints, the question is hence to know whether the subjects' behaviour in the verbal transformation tasks involved in Experiments 2 and 3 was actually driven mainly by articulatory coordination, or also by auditory templates. The following experiment was designed to test a possible perceptual alternative to the articulatory cohesion hypothesis by examining the persistence of verbal transformation asymmetries using a purely perceptual paradigm.

#### Method

##### *Participants*

Twenty-four students from Grenoble University participated in this experiment. All were native French speakers without hearing or speaking disorders, and were naive as to the purpose of the experiment.

##### *Phonetic material*

The /psə/, /səp/, /əps/, /spə/, /pəs/ and /əsp/ sequences were individually recorded into a digital audiotape by a trained phonetician (J.-L.S.) at a fixed speech rate of two cycles per second. The items were digitized to the hard disk of a PC computer at a sampling rate of 48-kHz with 16-bit quantization. Each sequence was then reduplicated one hundred times in an individual sound file with a 500 ms stimulus onset asynchrony [SOA].

---

---

Insert Figure 4 about here

---

---

### Apparatus

The stimuli were presented binaurally over headphones at a comfortable sound level. Transformations were collected via a microphone and directly recorded as individual sound files onto the hard disk of the computer.

### Procedure

The participants were tested individually in a quiet room. The experiment began with a lengthy briefing during which they were introduced with the verbal transformation task. Then they were told that they would hear an utterance being played repeatedly and were asked first to report what they hear and then to listen carefully for any changes in the repeated utterance. If the stimulus changed into another form, they were asked to report the transformation. It was indicated that the change could be subtle or very noticeable and could correspond to a word as well as a pseudoword. Finally, the participants were assured that there were no correct or incorrect responses and that if they did not hear a transformation, they were to say nothing. In the test session, the six stimuli were presented in one of twelve counterbalanced orders. Lengthy breaks were offered between trials.

### Results

The data were analyzed by labelling subjects' reports in the response sound files. Overall, 80.6% of the first reported forms matched the veridical repeated sequence. Furthermore, when the subject did not first report the correct sequence (e.g., /psəp/ instead of /psə/), 61% of the following transformation corresponded to the repeated utterance.

In the analyses presented below, only the transformations following a correct initial identification of the repeated utterances were taken into account. If the subject did not report a

transformation during the 30s. period after the first reported form, the sequence was considered as stable.

---



---

Insert Table 4 about here

---



---

On average, only 8% of the sequences remained stable and 3% were transformed according to a “shifting” parsing procedure (see Table 4a). The observed transformations from one group to the other were non-existent. Concerning the unpredicted transformations (on average 89% - see Table 4b), 22% of the overall responses involved lexical transformations (e.g., /sypɛr/ (“super”) for /spə/), 29% corresponded to a phonetic deviation (e.g., /sop/ for /səp/, /tse/ for /psə/), and 17% involved auditory streaming mechanisms (e.g., /əp/ for /əps/, /əs/ for /əsp/).

#### Stabilities and preferential transformations

Global statistical analyses yielded no significant effect of stimulus order ( $\chi^2(5) = 1.05$ ) and a significant global stimulus effect ( $\chi^2(5) = 15.02, p < .05$ ). Within each group, the statistical comparisons of stability and attractivity patterns showed the following results. Within Group 1, the global comparison of the observed stability per sequence was not significant ( $\chi^2(2) = 4.14$ ). Furthermore, none of the sequences were transformed according to a “shifting” parsing procedure. Within Group 2, the global comparison of the observed stability per sequence was significant ( $\chi^2(2) = 6.38, p < .05$ ), a result largely due to the greater stability of the /pəs/ syllable although none of the individual comparisons were significant. However, the global comparison of the observed attractivity per sequence was not significant ( $\chi^2(2) = 2.70, p < .005$ ).

#### Miscellaneous Transformations

According to the great number of transformations outside the two groups, we performed three distinct analyses across sequences related to their respective number of transformations based either on lexical transformation, phonetic deviation or auditory streaming. The results showed no discrepancies across sequences for the two first groups of transformations ( $\chi^2(5) = 9.32$ ,  $\chi^2(5) = 8.00$ , respectively) but a significant effect across sequences for the transformations involving an auditory streaming mechanism ( $\chi^2(5) = 41.80$ ,  $p < .001$ ), with the /əps/ and /əsp/ sequences showing a high number of transformations (62% and 32%, respectively).

### Discussion

Taken together, the stability and attractivity patterns of the sequences suggest that distinct constraints act on the elaboration of verbal representations during a perception procedure and a self-repetition procedure.

When compared to Experiment 2 and 3, all the sequences showed a lower degree of stability (on average, 8% vs. 51% and 52% respectively), a result in line with previous studies showing that perceptual stability constraints acting on verbal transformations are not fully equivalent for a perception procedure and a production variant. Lackner (1974) first reported that self-produced repetition of monosyllabic nonsense words resulted in far fewer speaker-perceived transformations than when the speaker's productions were played back to them. MacKay et al. (1993) further replicated this finding, showing that subjects experienced more transformations when they listened to a repeating word than when they overtly repeated the word. Altogether, these results suggest an increase of perceptual stability during an overt self-repetition mode. To explain this effect, Lackner proposed that perceptual mechanisms during self-repetition are alerted by a corollary discharge, or efference copy, that accompanies the motor execution of a

speech sequence (for an extent of the concept of efference copy to the on-line monitoring of one's own voice, see Frith, 1992). In the absence of such generated signal in the perception condition, informing on the forthcoming speech sound, the ability to maintain a stable perceptual representation should then decrease.

Another important result of Experiment 4 is the lower attractivity degree of the target sequences (on average 4% vs. 38% and 36% in Experiment 2 and 3). Indeed, contrary to Experiment 2 and 3 where transformations principally occurred within the same group of sequences, 89% of the present transformations corresponded to a lexical transformation, a phonetic deviation or were based on auditory streaming processes, three well-established transformation mechanisms during a perceptual procedure of the verbal transformation paradigm (e.g., Warren, 1961; Warren and Meyers, 1987; Kaminska, Pool and Mayer, 2000; Pitt and Shoaf, 2001, 2002; Shoaf and Pitt, 2002). Hence, the weak number of transformations relying on a shifting parsing process and altogether the completely different pattern of transformations compared with Experiments 2 and 3 clearly rules out a pure auditory interpretation of the verbal transformation asymmetries observed in the production conditions. This therefore reinforces the articulatory cohesion hypothesis as the most likely and coherent explanation of the asymmetries displayed in Experiments 2 and 3.

### GENERAL DISCUSSION

Investigate the causes of auditory illusions, which are viewed as windows into the linguistic processes that operate during veridical perception, provides a useful framework to enhance our understanding of the language system by revealing otherwise hidden mechanisms. From this point, the Verbal Transformation Effect appears to be well suited to examine how speech production and perception constraints may intervene on the emergence and analysis of verbal representations.

The aim of the present study was to explore whether specific speech production constraints, specifically temporal coherence between articulatory gestures, may intervene in the Verbal Transformation Effect. Having selected a convenient phonetic material and displayed variations in the temporal clustering of articulatory-acoustic events within this material (Experiment 1), we showed that these variations could lead to perceptual asymmetries in verbal transformations during an overt repetition procedure, thus suggesting that the perceptual stability and attractivity of an uttered sequence might also depend on articulatory cohesion constraints (Experiment 2). The fact that the same transformation trends were found during a covert repetition mode confirms some functional coupling between overt and covert speech and suggests that specific articulatory control constraints originating from the motor system may participate in the emergence of verbal representations in the human brain even without any external signal neither articulatory nor auditory (Experiment 3). Finally, the absence of such asymmetric bias in the transformations observed during a purely perceptual condition rules out a pure auditory interpretation of the verbal transformation asymmetries observed in the production conditions and shows that distinct constraints may act on the elaboration of verbal

representations during a perception procedure and a self-repetition procedure (Experiment 4).

#### Discarding pure lexical and phonological interpretations

One important source of influence in the verbal transformation paradigm comes from a set of general or language-specific linguistic constraints. In regard of our articulatory cohesion hypothesis, it is therefore important to check for a possible alternative explanation of the asymmetric bias observed in Experiment 2 and 3.

Firstly, verbal transformations should vary as a function of distinct lexical factors related to the repeated stimulus: the lexical status, the neighborhood density and the word frequency (see Table 5). Given that none of our speech stimuli occurs in the French lexicon, we consider that the lexical status cannot account for the stability and attractivity discrepancies between the speech sequences. Apart from the lexical status of the stimulus, it has been argued that a great number of neighbours (i.e., the number of lexical entries that are phonologically similar to the repeated stimulus; e.g., /pus/ ("thumb") for /pəs/) should increase the number of possible primed lexical candidates, resulting in a greater number and a wider range of transformations and then entailing a lower stability of the stimulus (Yin and MacKay, 1992; MacKay et al., 1993). However, this neighbourhood density effect does not appear in accordance with our results. Indeed, the weak stability for the /əps/ and /əsp/ sequences observed, once the glottal stop effect taken into account in Experiment 2, cannot be explained by their respective neighbourhood density values, these sequences showing a weaker (or at least a similar) number of neighbours compared to the other sequences. Hence, neighborhood density is not likely to provide the explanation of the present asymmetries. Another concurrent interpretation could come from the

lexical frequency of the speech sequences in the participants' lexicon. Indeed, this lexical frequency could bias transformations towards a sequence with a greater number of lexical entries (MacKay et al., 1993). Considering the results of Experiment 2 and 3, it appears that, in Group 1, this lexical frequency effect is not in accordance with the stronger attractivity of /psə/, /səp/ being the most favored sequence in terms of lexical entries. However, in Group 2, there seems to be a slight advantage of /pəs/ over /spə/ in terms of attractivity, though it was always below the significance threshold. This trend, not contained in our predictions, could be due to the fact that /pəs/ displays a greater number of lexical entries than /spə/, thus underlining a potential effect of the lexical frequency during the task. This position is reinforced considering the results of Experiment 4, showing a greater attractivity of /pəs/, although not significant, as compared to the others sequences.

---



---

Insert Table 5 about here

---



---

Another hypothesis based on phonetic or phonotactic regularities could be that preferential transformations derive from syllabic structure constraints. The study of typological trends in syllable structure (see Table 6 for an overview of syllabic structure frequencies extracted from a sample of geographically and genetically dispersed languages of the ULSID database; Maddieson, 1984; Vallée et al., 2000) shows that VCC and CCV syllables are very infrequent in phonological inventories and that CVC is the most frequent syllable after CV. The results of Experiment 2 and 3 seem indeed to satisfy the largely shared constraint of avoidance of syllables with no consonantal onset: CVC and CCV syllables do not switch towards VCC, just as

CV did not switch towards VC (Stetson, 1951; Tuller and Kelso, 1990, 1991; de Jong, 2001; de Jong, Nagao and Lim, 2002). However, the large number of transformations from CVC /səp/ to CCV /psə/ in Group 1, which does violate the constraint of CVC stability, allows us to discard this alternative interpretation based only on syllabic structure regularities.

---



---

Insert Table 6 about here

---



---

Hence, none of these linguistic factors can fully explain the observed patterns of both stability and attractivity in our experiments. A last concurrent interpretation must however be considered quite seriously. Indeed, our results appear compatible with several phonological theories of syllabification (for a review, see Goslin and Frauenfelder, 2000) and are relevant to the syllabic segmentation issue in psycholinguistics (e.g., Treiman and Danis, 1988; Content, Kearns, and Frauenfelder; 2001; Dumay, Frauenfelder and Content; 2002). Whatever the group, the observed pattern of attractivity respected the Sonority Sequencing Principle (Clements, 1990). According to this theory, a preferred syllable shows a sonority profile (or sonority scale) that maximally rises towards the nucleus peak and minimally falls towards the end of the syllable. This sonority principle, together with the preference for onsets over codas according to the Obligatory Onset Principle (Hooper, 1972) and the Maximal Onset Principle (Pulgram, 1970), would predict hierarchies like "/psə/ > /səp/ > /əps/" in Group 1 and "/pəs/ > /spə/ > /əsp/" in Group 2, more or less compatible with the observed stability and attractivity patterns. In this respect, the present results might be considered as relevant to the syllabic segmentation issue. Particularly, the fact that the patterns of preferential transformations are maintained in

covert speech brings an important indication about the ability of syllabification mechanisms to intervene in the speaker's brain.

However, the sole explanation based on syllabification theories does not fully account for the observed results. Indeed, syllabification rules should lead to within-groups hierarchies that are not in agreement with our data concerning the non significant /pəs/ vs. /spə/ difference. Nor can they explain the pervasive trend in this study to shift from Group 2 to Group 1, especially towards the sequence /psə/ (see the many intrusions of /psə/ transformations for most stimuli in Tables 2a and 3a), which happens to be the best sequence in terms of articulatory coherence in our predictions (Table 1). Moreover, combining syllabification mechanisms, syllabic structure constraints and lexical factors should lead to a weak preference for /psə/ over /səp/ (/səp/ appears more frequently in the participants' lexicon than /psə/ and corresponds to a "good" CVC syllable in terms of sonority and syllabic structure) and to a large preference for /pəs/ over /spə/ (/pəs/ being both preferred in terms of syllabic structure, sonority sequencing and lexical frequency). This is obviously different from the obtained results.

Therefore, the predicted articulatory cohesion scale – which is of course related in some sense with syllabification principles – seems to provide the most likely and coherent explanation of the results of Experiment 2 and 3. Thus, the fact that neither universal nor language specific constraints can fully account for the data – in particular, for the success of the /psə/ sequence – argues in favour of the existence of specific articulatory control constraints acting on verbal transformations during a self-repetition mode.

Relations between speech perception and production in the Verbal Transformation Effect

A major result of the present set of experiments concerns the completely different pattern of verbal transformations obtained in the production procedure in Experiment 2, or its covert variant in Experiment 3, vs. the perception procedure in Experiment 4. There are two major differences. Firstly, the *number* of transformations is much larger in Experiment 4, in agreement with previous experiments (Lackner, 1974; MacKay, 1993). This has been proposed as due to a corollary discharge mechanism (Lackner, 1974) or a top-down priming process (MacKay, 1993) stabilizing the speech representation in the production procedure. Secondly, the *pattern* of transformations is also completely different, and this is clearly a new finding. Actually, some of the transformations are shared in both procedures, considering the unexpected transformations in Experiment 2 and 3 which provide a number of previously underlined contents including substitution of a phoneme by a phonetically close one (Warren, 1961; Warren and Meyers, 1987), auditory streaming (Pitt and Shoaf, 2001, 2002) and lexical transformations (Warren, 1961; Kaminska, Pool and Mayer, 2000, Shoaf and Pitt, 2002). However, while these transformations mechanisms were the majority in the pure auditory procedure, it was not the case during the self-repetition conditions in which most of the transformations depended on a shifting parsing process driven by the degree of synchronization between articulatory gestures. In these latter experiments, the more temporally clustered sequences played the role of attractors in verbal transformations (in particular, the "all-phased" monostable /psə/). Taken together, these results suggest that transformation mechanisms are not fully equivalent for a perception procedure and a production variant one, with articulatory control constraints as a major factor in transformations during a self-repetition procedure.

Although our results seem to point out articulatory control constraints as the major factor in transformations during the production experiments, it could however be proposed that multisensory representations, combining auditory and proprioceptive components, drove the search for temporal clustering in Experiment 2 and Experiment 3 (in this latter case, multisensory imagery produced by the “inner voice” would play the same role). Indeed, the speakers in Experiment 2 both *produce* gestures and *perceive* them through various sensory channels, and it is quite likely that both motor and perceptual requirements shape their behaviour. The same is true in Experiment 3, where the perceptuo-motor loop between the “inner voice” and the “inner ear” is involved in the brain in a covert mode. This is even clearer considering a recent functional brain imaging study carried out in our laboratory (Sato et al., 2004). In this fMRI experiment, two conditions were contrasted: a baseline condition involving the simple mental repetition of the speech sequences used in the present study and a verbal transformation condition involving the mental repetition of the same items with an active search for verbal transformation. The contrast between the verbal transformation task and the baseline revealed a left-lateralized network of activations, notably within the inferior frontal gyrus, the supramarginal gyrus and the superior temporal gyrus – three areas considered to be involved in the analysis of articulatory-based representations, in the interfacing between sound-based and articulatory-based representations of speech, and in the self-monitoring of verbal material, respectively. These results thus strongly suggest that the Verbal Transformation Effect shares common components of speech perception and speech production and that it relies both on sound-based and on articulatory-based representations. On the other hand, the present results underlined the fact that transformation mechanisms do not act with the same extent for a

perception procedure and a production variant one, with articulatory control constraints as a major factor in transformations during a self-repetition procedure.

In conclusion, the set of experiments presented in the present study demonstrate in a coherent way that the perceptual stability and attractivity of an uttered sequence depend on articulatory control constraints which are hence likely to be involved, together with auditory, phonological and lexical constraints, in the emergence and analysis of verbal representations in the human brain.

ACKNOWLEDGMENTS

We thank Mark Pitt, Kenneth de Jong, and three anonymous reviewers for an helpful review of the manuscript. We also thank Alain Content, Pascal Perrier, Daniel Reisberg and Rudolph Sock for helpful discussions and Anahita Basirat, Anne Vilain, Christophe Savariaux and Willy Serniclaes for their help in scoring and analysing data. Aspects of this paper were presented at the XV<sup>th</sup> International Congress of Phonetic Sciences, 2003, Barcelona.

REFERENCES

- Abry, C., Cathiard, M.-A., Vilain, A., Laboissière, R. & Schwartz, J.-L. (2004). Some insights in bimodal perception given for free by the natural time course of speech production. In: G. Bailly, P. Perrier & E. Vatikiotis-Bateson (Eds) *Festschrift Christian Benoît*. MIT Press, Cambridge.
- Browman, C.P. & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6: 201-251.
- Chambers, D. & Reisberg, D. (1985). Can mental images be ambiguous? *Journal of Experimental Psychology: General*, 11: 317-328.
- Clements, G.N. (1990). The role of the sonority cycle in core syllabification. In M.E. Beckman & J. Kingston (Eds), *Papers in Laboratory Phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press, pp. 283-333.
- Content, A., Kearns, K.R. & Frauenfelder, U.H. (2001). Boundaries versus onsets in syllabic segmentation. *Journal of Memory and Language*, 45: 177-199.
- de Jong, K.J. (2001). Rate-induced resyllabification revisited. *Language and Speech*, 44: 197-216.
- de Jong, K.J., Nagao, K. & Lim, B.J. (2002). The interaction of syllabification and voicing perception in American English. *ZAS Papers in Linguistics*, 28: 27-38.
- Dufour, S., Peereman, R., Pallier, C. & Radeau, M. (2002). VoCoLex: A lexical database on phonological similarity between French words. *L'Année Psychologique*, 102: 725-746.
- Dumay, N., Frauenfelder, U.H. & Content, A. (2002). The role of the syllable in lexical segmentation in French: Word-spotting data. *Brain and Language*, 81: 144-161.

- Frith, C.D. (1992). *The Cognitive Neuropsychology of Schizophrenia*. Hove, Lawrence Erlbaum.
- Gleason, P., Tuller, B. & Kelso, J.A.S. (1996). Syllable affiliation of consonant clusters undergoes a phase transition over speaking rates. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, Vol. 1, pp. 276-278.
- Goslin, J. & Frauenfelder, U.H. (2000). A comparison of theoretical and human syllabification. *Language and Speech*, 44(4): 409-436.
- Hooper, J.B. (1972). The syllable in phonological theory. *Language*, 48: 525-540.
- Jakobson, R. (1966). Implications of language universals for linguistics. In J.H. Greenberg (Ed.), *Universal of language*, MIT Press, Cambridge: 263-278.
- Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Science*, 17: 187-245.
- Kaminska, Z., Pool, M. & Mayer, P. (2000). Verbal transformation: Habituation or spreading activation? *Brain and Language*, 71: 285-298.
- Lackner, J.R. (1974). Speech production: evidence for corollary-discharge stabilization of perceptual mechanisms. *Perceptual and Motor Skills*, 39: 899-902.
- Lallouache, M.T. (1990). Un poste 'visage-parole'. Acquisition et traitement de contours labiaux. *Actes des XVIIIèmes Journées d'Études sur la Parole*, Montréal, pp. 282-286.
- Landauer, T.K. (1962). Rate of implicit speech. *Perceptual and Motor Skills*, 15: 646.
- Luce, P.A., Pisoni, D.B. & Goldinger, S.D. (1990). Similarity neighborhoods of spoken words. In G.T.M. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MIT Press, pp. 122-147.
- MacKay, D.G. (1982). The problems of flexibility, fluency, and speed-accuracy trade-off in

- skilled behavior. *Psychological Review*, 89: 483-527.
- MacKay, D.G., Wulf, G., Yin, C. & Abrams, L. (1993). Relations between word perception and production: New theory and data on the verbal transformation effect. *Journal of Memory and Language*, 32: 624-646.
- MacNeilage, P.F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21: 499-511.
- MacNeilage, P.F. & Davis, B.L. (2000). Origin of the internal structure of words. *Science*, 288: 527-531.
- Maddieson, I. (1984). *Patterns of Sounds*. Cambridge University Press, Cambridge.
- Natsoulas, T. (1965). A study of the verbal transformation effect. *American Journal of Psychology*, 78: 257-263.
- Postma, A. & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech*, 39(4): 375-392.
- Pitt, M. & Shoaf, L. (2001). The source of a lexical bias in the Verbal Transformation Effect. *Language and Cognitive Processes*, 16(5/6): 715-721.
- Pitt, M. & Shoaf, L. (2002). Linking verbal transformations to their causes. *Journal of Experimental Psychology: Human Perception and Performance*, 28: 150-162.
- Pulgram, E. (1970). *Syllable, Word, Nexus, Cursus*. The Hague: Mouton.
- Reisberg, D. (1992). *Auditory Imagery*. Lawrence Erlbaum, Hillsdale.
- Reisberg, D., Smith, J.D., Baxter, A.D. & Sonenshine, M. (1989). "Enacted" auditory images are ambiguous; "pure" auditory images are not. *Quarterly Journal of Experimental*

- Psychology*, 41A: 619-641.
- Sato, M. & Schwartz, J.-L. (2003). Linking speech, verbal imagery and working memory: Articulatory control constraints in the verbal transformation effect. *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelona, pp. 435-438.
- Sato, M., Baciú, M., Lœvenbruck, H., Schwartz, J.-L., Cathiard, M.-A., Segebarth, C. & Abry, C. (2004). Multistable representation of speech forms: An fMRI study of verbal transformations. *NeuroImage*, 23(3): 1143-1151.
- Smith, J.D., Reisberg, D. & Wilson, M. (1995). The role of subvocalization in auditory imagery. *Neuropsychologia*, 11: 1433-1454.
- Shoaf, L. & Pitt, M. (2002). Does node stability underlie the verbal transformation effect? A test of node structure theory. *Perception & Psychophysics*, 64(5): 795-803.
- Stetson, R.H. (1951). *Motor Phonetics: A study of speech movements in action*. Amsterdam: North-Holland.
- Treiman, R. (1983). The structure of spoken syllables: Evidence from novel word games. *Cognition*, 15: 49-74.
- Treiman, R. & Danis, C. (1988). Syllabification of intervocalic consonants. *Journal of Memory and Language*, 27: 87-104.
- Tuller, B. & Kelso, J.A.S. (1990). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and Performance XIII*, Hillsdale, NJ: Erlbaum: 429-452.
- Tuller, B. & Kelso, J.A.S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research*, 34: 501-508.

- Vallée, N., Boë, L.J., Maddieson, I. & Rousset, I. (2000). Des lexiques aux syllabes des langues du monde – Typologie et structures. *Actes des XXIIIèmes Journées d'Etude sur la Parole*, Aussois, pp. 93-96.
- Warren, M.R. & Gregory, R.L (1958). An auditory analogue of the visual reversible figure. *American Journal of Psychology*, 71: 612-613.
- Warren, M.R. (1961). Illusory changes of distinct speech upon repetition – The verbal transformation effect. *British Journal of Psychology*, 52: 249-258.
- Warren, M.R. (1982). *Auditory Perception*. Pergamon Press, New-York.
- Warren, M.R. & Meyers, D.M. (1987). Effects of listening to repeated syllables: Category boundary shifts versus verbal transformation. *Journal of Phonetics*, 15: 169-181.
- Yin, C. & MacKay, D.G. (1992). Auditory illusions and aging: Transmission of priming in the verbal transformation paradigm. Paper presented to the *IVth Biennial Cognitive Aging Conference*, Atlanta.

FIGURE LEGENDS

Figure 1. Unfolding of acoustic events for the /psə/, /səp/, /əps/, /pəs/, /spə/ and /əsp/ speech sequences. FO: Frication onset for [s], VO: Voicing onset for [ə], BR: Bilabial release for [p] after the bilabial closure. For each sequence, the acoustic signal (top) is displayed in synchrony with the corresponding spectrogram (temporal frequency representation, bottom).

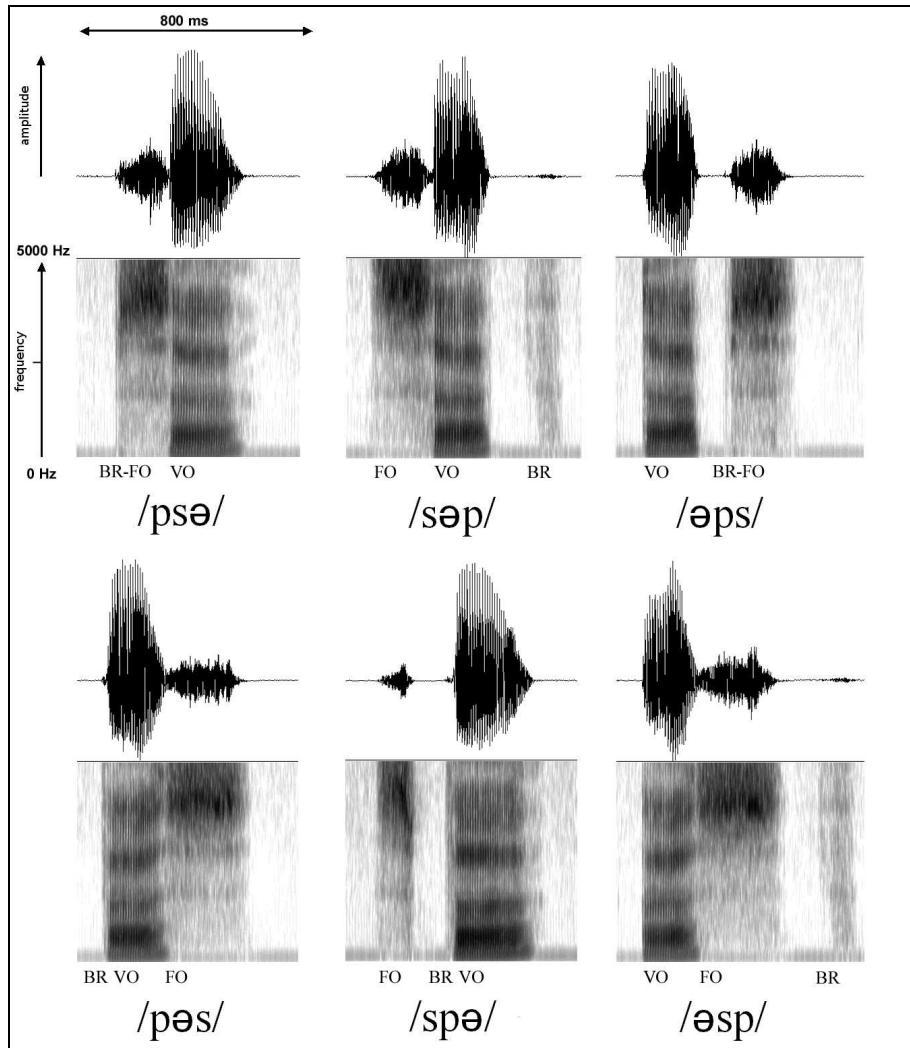


Figure 2.1. Variations of the coherence index of articulatory-acoustic events (defined as the time separating the onsets of [p] and [s] divided by the cycle duration) as a function of cycle duration (in sec) for the /psə/ and /səp/ speech sequences in the increased rate condition.

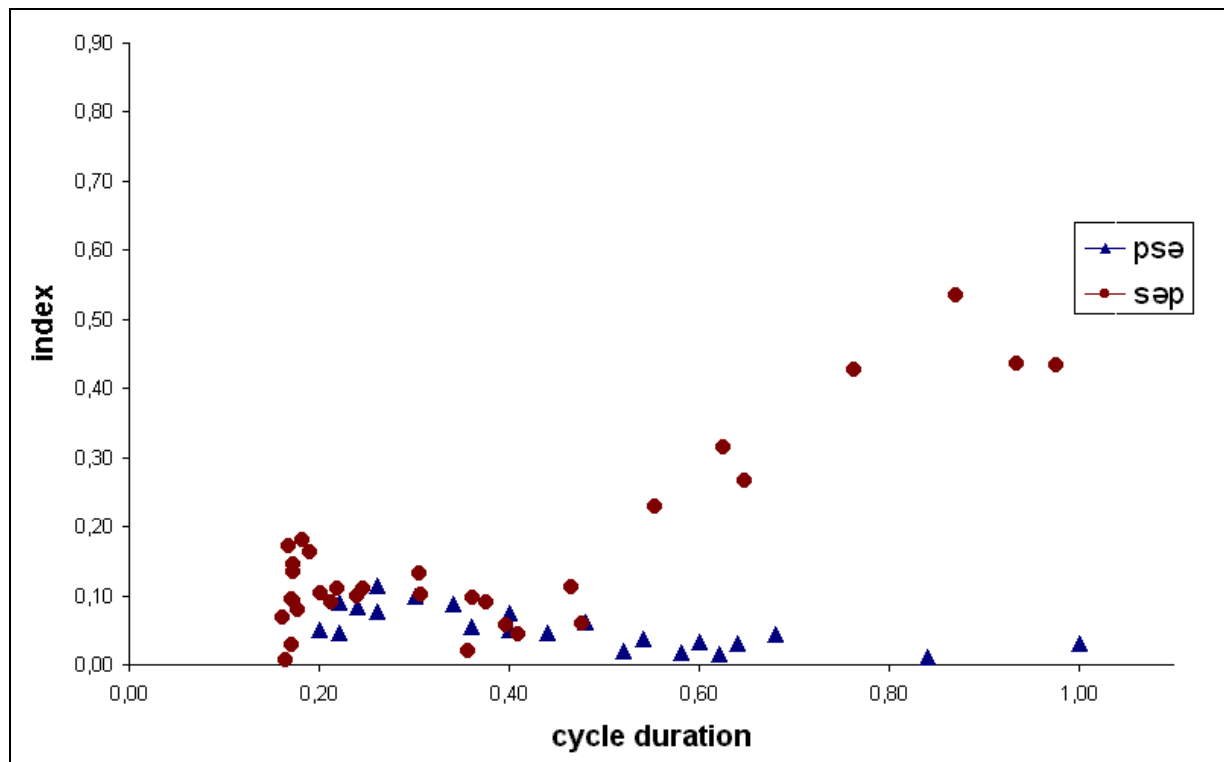


Figure 2.2. A. Variations of the coherence index of articulatory-acoustic events (defined as the time separating the onsets of [p] and [s] gestures divided by the total cycle duration) as a function of cycle duration (in sec) for the /psə/ and /səp/ speech sequences in the fixed rate condition. B. Variations of the mean index (with standard deviation) according to a low (around 1 cycle per second) or moderate (around 2 cycles per second) speaking rate.

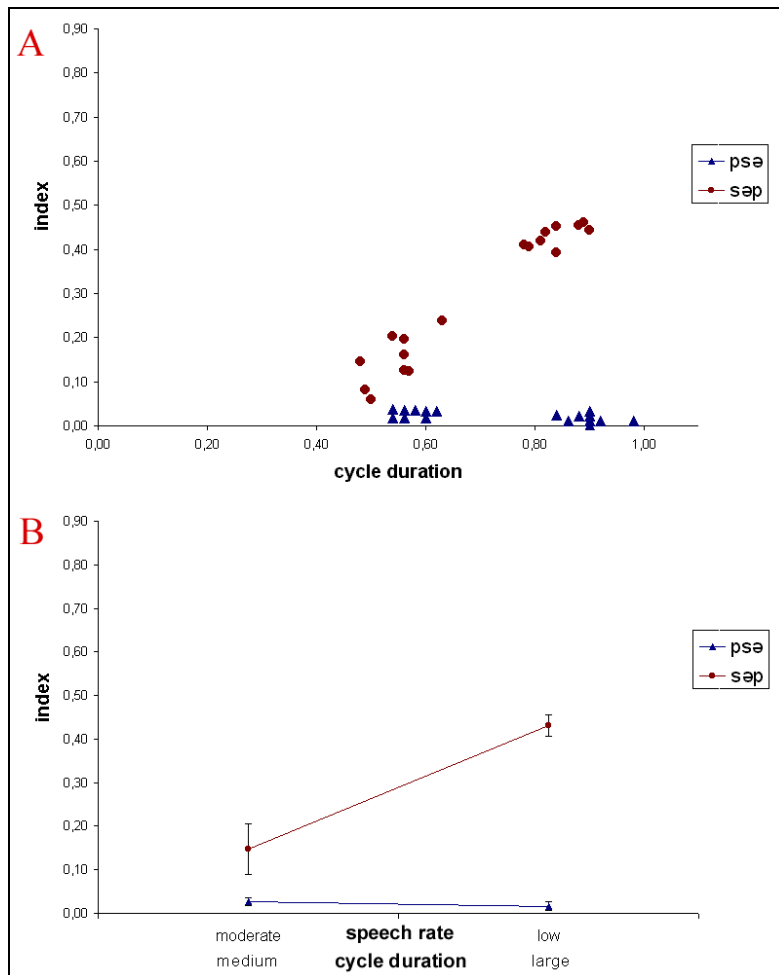


Figure 3.1. Variations of the coherence index of articulatory-acoustic events (defined as the time separating the onsets of [p] and [s] divided by the total cycle duration) as a function of cycle duration (in sec) for the /pəs/ and /spə/ speech sequences in the increased rate condition.

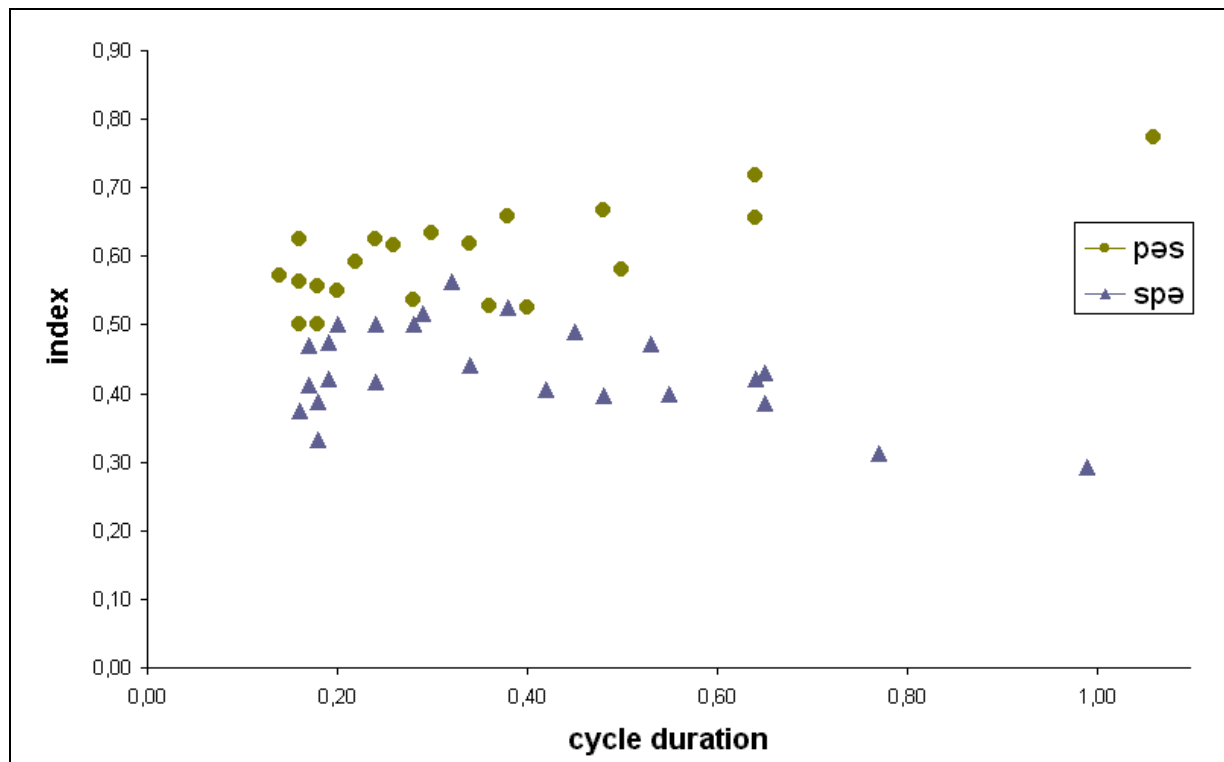
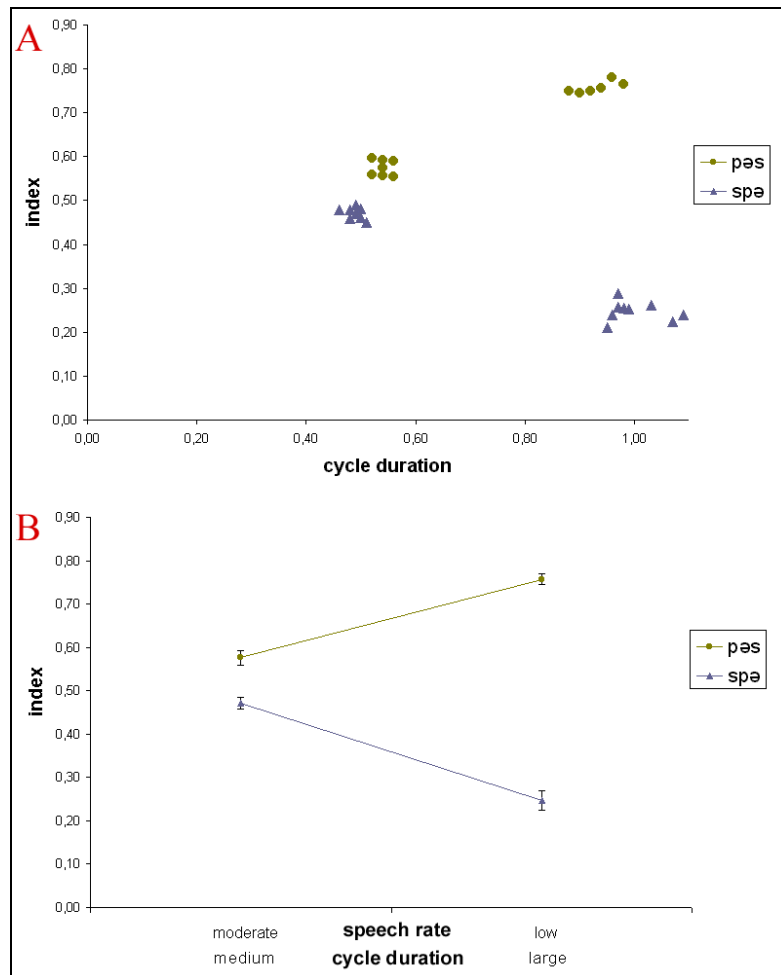


Figure 3.2. A. Variations of the coherence index of articulatory-acoustic events (defined as the time separating the onsets of [p] and [s] divided by the total cycle duration) as a function of cycle duration (in sec) for the /pəs/ and /spə/ speech sequences in the fixed rate condition. B. Variations of the mean index (with standard deviation) according to a low (around 1 cycle per second) or moderate (around 2 cycles per second) speaking rate.



TABLES

Table 1. Sequence classification according to the degree of articulatory cohesion between the consonantal and vocalic gestures and expected transformations during the repetition process.

Sequence	Degree of Articulatory Cohesion	Prediction
/psə/	Strong – onset cluster and vowel synchronized, consonants in the onset synchronized	/psə/
/səp/	Average – onset consonant and vowel synchronized, coda desynchronized	/psə/
/əps/	Weak – vowel and consonant cluster desynchronized, consonants in the coda synchronized	/psə/
/pəs/	Average – onset consonant and vowel synchronized, coda desynchronized	/pəs/ - /spə/
/spə/	Average – onset cluster and vowel synchronized, consonants in the onset desynchronized	/pəs/ - /spə/
/əsp/	Very weak – vowel and consonantal cluster desynchronized, consonants in the coda desynchronized	/pəs/ - /spə/

**Table 2.** (A) Transformation frequencies observed in Experiment 2. Stimulus sequences in rows, transformations in columns (N=56; “Misc” = Miscellaneous transformations). (B) and (C) Stability and weighted attractivity degrees per sequence within Group 1 and within Group 2 (the sequences predicted as the more stable and attractive are underlined).

(A)

Sequence	⇒ /psə/	⇒ /səp/	⇒ /əps/	⇒ /pəs/	⇒ /spə/	⇒ /əsp/	⇒ misc.
/psə/	.75	.18			.02		.05
/səp/	.50	.30	.02				.18
/əps/	.29		.64				.07
/pəs/				.41	.43		.16
/spə/	.04			.46	.39		.11
/əsp/	.07		.04	.20	.05	.55	.09

(B)

Sequence	Stability	Attractivity
<u>/psə/</u>	.75	.75
/səp/	.30	.29
/əps/	.64	.02

(C)

Sequence	Stability	Attractivity
<u>/pəs/</u>	.41	.63
<u>/spə/</u>	.39	.47
/əsp/	.55	.00

**Table 3.** (A) Transformation frequencies observed in Experiment 3. Stimulus sequences in rows, transformations in columns (N=29; “Misc” = Miscellaneous transformations). (B) and (C) Stability and weighted attractivity degrees per sequence within Group 1 and within Group 2 (the sequences predicted as the more stable and attractive are represented in bold).

(A)

Sequence	⇒ /psə/	⇒ /səp/	⇒ /əps/	⇒ /pəs/	⇒ /spə/	⇒ /əsp/	⇒ misc.
/psə/	.63	.10		.07	.10		.10
/səp/	.31	.62					.07
/əps/	.42	.07	.34				.17
/pəs/	.07			.69	.21		.03
/spə/	.10	.03		.11	.48		.28
/əsp/	.21	.03	.07	.17	.07	.35	.10

(B)

Sequence	Stability	Attractivity
<b>/psə/</b>	.63	.70
/səp/	.62	.17
/əps/	.34	.00

(C)

Sequence	Stability	Attractivity
<b>/pəs/</b>	.69	.24
<b>/spə/</b>	.48	.29
/əsp/	.35	.00

**Table 4.** (A) Transformation frequencies observed in Experiment 4. Stimulus sequences in rows, transformations in columns (N=24; “Misc” = Miscellaneous transformations). (B) Description of the miscellaneous transformations.

(A)

Sequence	⇒ /psə/	⇒ /səp/	⇒ /əps/	⇒ /pəs/	⇒ /spə/	⇒ /əsp/	⇒ misc.
/psə/							1.00
/səp/		.10					.90
/əps/							1.00
/pəs/				.29	.06		.65
/spə/				.05	.05		.90
/əsp/				.11		.05	.84

(B)

Sequence	⇒ Lexical Transformation	⇒ Auditory Streaming	⇒ Phonetic Deviation	⇒ Other Transformations
/psə/	.26		.53	.21
/səp/	.35		.45	.10
/əps/		.62	.19	.19

---

/pəʃ/	.12	.06	.24	.23
/spə/	.35		.20	.35
/əʃp/	.21	.32	.16	.15

---

Table 5. Lexical type frequency (LTF; defined as the number of lexical entries incorporating a monosyllabic structure identical to that of the stimulus at any position in a word) and the neighborhood density (ND; defined as the number of phonologically similar words that differ from the stimulus only by a single substitution, insertion or deletion at any position in the target word; Luce, Pisoni and Goldinger, 1990). For each of the measures, the sum (stf) and range of associated token frequencies are indicated (in number of occurrences per million). All lexical analyses were extracted from VoCoLex, a lexical database for the French language (~105000 words; Dufour, Peerman, Pallier and Radeau, 2002).

Sequence	L.T.F	stf	range	N.D	stf	range
/psV/	114	81	0-13	31	11357	0-5031
/sVp/	371	812	0-92	59	20310	0-5031
/Vps/	131	170	0-118	34	3206	0-2096
/spV/	674	1248	0-119	19	10160	0-5031
/pVs/	1379	10676	0-6372	118	54538	0-16011
/Vsp/	598	1539	0-229	14	9320	0-8743

Table 6: Percentage of syllabic structures in the ULSID database (UCLA Lexical and Syllabic Inventory Database). From Vallée et al. (2000).

Type	Percentage	Type	Percentage
CV	.545	CCVC	.013
CVC	.362	CVCC	.006
V	.044	CCV	.005
VC	.025	VCC	.000