The Syllable's Role in Speech Segmentation

Jacques Mehler, Jean Yves Dommergues, and Uli Frauenfelder
CNRS

AND

Juan Segui
Universite Rene Descartes et EPHE associe au CNRS

In this study a monitoring technique was employed to examine the role of the syllable in speech segmentation. The perceptual unit into which the continuous speech signal is segmented constitutes one of the central issues in psycholinguistics. Investigators have argued for the perceptual primacy of a wide variety of candidates such as the phoneme, syllable, word, and sentoid. Traditionally, the phoneme has been taken as the basic perceptual unit. However, the lack of invariance between the acoustic signal and the phoneme has weakened its candidacy. Indeed, the acoustic information corresponding to a phoneme is often distributed throughout several segments or inversely, the information about a phoneme is concentrated into one short stretch of acoustic signal (Liberman, 1970).

Such facts as these have led some to abandon the phoneme and endow the syllable with perceptual primacy. The feasibility of this view is corroborated by the incorporation of the syllable into several models of speech recognition systems (Mermelstein, 1980). Further evidence is given by the fact that the syllable has also accumulated a large number of domains. By means of a recognition technique, Massaro (1974) has shown that the existence of a 250-millisecond storage period in which a phoneme is perceived is not necessary for perception of the syllable.

There are also some data that classify the phonological status of the syllable as a basic perceptual unit. Liberman, Weiler, Fischer, and Carter (1972) tested young children on a task in which they had to identify the syllable in a target. They found that children of different ages could identify syllables, but not phonemes. Segui (1980) has shown that older children (6-7 years old) perform better on this task than younger children (4-5 years old).

The authors wish to express their thanks to Yves Barbini for his invaluable technical assistance. This work was supported by grants from Harry Frank Guggenheim Foundation 1977/1978 and D.O.R.S.T. 1980 to J. Mehler. Requests for reprints should be sent to Dr. J. Mehler, Laboratoire de Psychologie, CNRS, 54, Boulevard Raspail, 75006 Paris.
phoneme based predictions can be made by hypothesizing a mechanism not based on syllabic structure but on the size of the target. First, if the subject conducts a phoneme-by-phoneme analysis of the stimulus, shorter RTs could be expected for the shorter target sequences (RT: $pa < paI$) since less of the stimulus word (one fewer phoneme) must be analyzed to initiate a response. Finally the inverse results (RT: $pa > pal$) is predicted by the uncertainty hypothesis (Foss & Swinney, 1973; Swinney & Pater, 1980) according to which RTs depend on the uncertainty of the subjects concerning the nature of the target/ stimulus. The smaller the target (in number of phonemes?), the fewer cues there are available for identifying the stimulus and hence the longer the RTs.

These three hypotheses will be put to test in the following experiment.

**EXPERIMENT I**

**Method**

*Subjects.* Forty-two subjects (divided into two groups of 21 each), all native French speakers from the Parisian university community, participated in the experiment which lasted about one half hour.

*Materials and design.* Five pairs of monomorphemic bisyllabic French nouns of similar frequency sharing the same initial three phonemes (CVC) were selected such that these phonemes made up the first syllable for one member of the pair and the first two (CV) phonemes formed the first syllable of the second member of the pair. For instance, in the pair: *palais*/*palmier* the first three phonemes (/pa/l/lu/) are identical. Yet, this CVC sequence corresponds to the first syllable only the word *palmier*, CV being the first syllable of the word *palais*. For each of the five pairs the initial consonant was either a voiced or a voiceless stop and the second was a liquid (either /l/ or /r/). The five pairs were: *palais*/*palmier*, *carotte*/*carotte*, *tarif*/*tarine*, *garage*/*garage*, *service*/*service*.

These 10 stimulus words containing the target were each placed as the last item in experimental sequences. Each experimental sequence was composed of the target item and 1 to 4 bisyllabic filler words. Both members of a pair appeared in the same position (from the second to the fifth position) in their respective sequences. In addition to this first set of 10 target sequences, a second set was used with different fillers but with the same target words in the same position as in the first set. Each set of target sequences was mixed with 10 different distractor sequences forming two blocks (A & B), each containing 20 sequences. Distractor sequences either had target words as the last item (any position from the first to the sixth) or no target to prevent the subjects from anticipating last items in long sequences. These 40 sequences along with 5 warm-ups were recorded by a French native speaker at a normal rate with a two-track Ampex AG440B. The words in each sequence were separated by 2-second intervals; sequences were separated by 10 seconds.

Table I provides a description of the experimental sequences and of the targets given to the two groups of subjects. Two groups were used to counterbalance the presentation order of a given stimulus word and its two corresponding targets.

As can be seen in Table 1, the stimulus word *balance* in position 2 is associated with the target *bal* for group 1 and with *bal* for group 2 in sequence number 6 of Block A. The inverse matching is found in sequence number 19 of Block B. The targets to be detected were displayed visually on small $3 \times 5$ inch cards numbered from 1 to 45. Before hearing each sequence, subjects heard the instruction "next card" and had 10 seconds to turn and read the next card. Subjects in groups 1 and 2 received different decks of cards.

The list was presented binaurally and subjects responded by pressing a response button that stopped an electronic clock in a PDP-12 computer. A click, aligned manually with the beginning of the target word, triggered the clock. A correction for each click was obtained by means of a two channel oscilloscope and was edited into the data collection program.

**Results**

The mean reaction times for each subject and for each item were computed. RTs longer than 1000 milliseconds and shorter than 100 milliseconds were omitted from the calculation of the means. The results obtained in the table are displayed in Figure 1. They show that when the subjects respond to a CV target in a word structure is CV/... (here to as CV words) their RTs are when they respond to the same CVC target (352 msec vs 371 msec).
When responding to a CVC word with a CVC target, subjects respond faster than when responding to the same word with a CV target (356 msec vs 378 msec, respectively). Since the results for Group I and Group 2 are analogous, a two-way within-subject analysis of variance was conducted using the combined data. The two main factors were target type (CV vs CVC) and word type (CV words vs CVC words). The interaction between these factors is highly significant, $F(1,42) = 11.03, p < .005$, but individually they failed to yield any significant difference, $F < 1$ in both cases. Likewise, a two-way analysis of variance was carried out on the mean reaction times corresponding to the five pairs of words giving $F(1,4) = 6.75, p < .10$ for the interaction between the two main factors. As before, individually these factors were not significant.

A $t$ test was used to compare the mean RTs for both word types as a function of the target stimulus. These results are presented in Table 2. The presence of a significant difference was obtained for both word types; for ev words, $t(4) = 2.67, p < .05$, for CVC words, $t(4) = 2.59, p < .01$. A $t$ test using items as the random variable gave analogous differences for CV words, $t(4) = 2.67, p < .05$, for CVC words, $t(4) = 2.36, p < .05$.

**Discussion**

These results show that subjects detect a target phoneme sequence faster when it corresponds to the first syllable of the stimulus word than when it does not. Thus, the RTs do not depend on target size as predicted by the two alternative hypotheses but rather on the syllabic relationship between the target and the stimulus word. The following (albeit simplistic and schematic) description of what subjects are likely to do in this task provides insight into these results. First, upon seeing the target, subjects must form and store some representation of it. On subsequently hearing the acoustic stimulus, they must segment and analyze it to produce a stimulus representation. The detection process can thus be seen as the matching of stimulus and target representations.

The analysis of the acoustic signal corresponding to a given target word (palmer) is unlikely to vary as a function of the target (pal, /ipal/) given to the subject. The observed RT differences can then be attributed to the process of matching between the stimulus and target representations and not to the computation of the former. When the target sequence corresponds exactly to the first syllable of the stimulus word, there is a better match between the two representations than when it does not. The varying degree of compatibility between the target and the stimulus representation could thus explain the direction of the RT differences. Since the representation of a given target (pa) presumably does not vary, the processing of the two words in a pair (palmer, palace) must differ. Thus there must be acoustic cues in the signal that the subjects exploit to derive different stimulus representations for the two words. This interpretation can explain the highly significant interaction found between word and target types.

This account is compatible with the target–stimulus mismatch hypothesis proposed by Mills (1980a, 1980b). Mills argues that the closer the match is between the subjects' expectancy about the stimulus and the actual acoustic stimulus, the faster the subjects respond to the stimulus. It should be noted, however, that this explanation cannot account for the RT differences found here without appealing to a syllabic segmentation unit. Indeed, a target–stimulus mismatch hypothesis, formulated in phonemic terms would not predict the observed interaction. If the subjects' uncertainty concerning the stimulus is based only on the number of phonemes in the target, then longer targets should have led to shorter RTs. Thus, a syllable-based hypothesis is necessary. Unfortunately, however, our experiment does not allow us to make precise claims concerning the target representation and the more-or-less analyzed character of the syllable.

**Experiment II**

**Method**

Subjects. Twenty subjects from the Parisian university community participated in this experiment.

Materials and design. The experimental design and the linguistic materials were the same as those used in the first experiment. Two groups of 10 subjects received the same instructions with the exception that they were told that the visually specified target corresponded to the vowel in the first syllable and the VC target to the first VS sequence in the word.

**Results**

The mean reaction times for each subject and for each item were computed. Reaction times longer than 1500 milliseconds and shorter than 100 milliseconds were omitted from these calculations. The excluded data made up 6.5% of the subjects' responses. Table 2 shows the overall reaction times.

<table>
<thead>
<tr>
<th>Target type</th>
<th>CV word</th>
<th>VC word</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>615</td>
<td>704</td>
</tr>
<tr>
<td>VC</td>
<td>695</td>
<td>710</td>
</tr>
</tbody>
</table>

**Discussion**

The results observed for VC targets port our hypothesis, according to which RTs are expected when the sequence belongs to one syllable rather than to different syllables. Further differences were obtained for the CV words. A syllabic segmentation hypothesis is unlikely to vary as a function of the target stimulus word, when it is contained in the same syllable (palmer) than when it is found in two different syllables (palace). This prediction is derived from the general hypothesis according to which subjects respond faster to stimulus items when these belong to the same rather than to different constituents at any linguistic level (Fodor, Bever, & Garrett, 1974).

**General Discussion**

Results of experiments I and II are at first sight not particularly compatible with the general hypothesis that syllable constitutes a unit of speech perception. Indeed, through a process of segmentation, subjects seem able to identify syllable-like constituents for the p words and sentences. However, the major problems with a syllable hypothesis was raised by Liberman and Kay (1978). These authors noted that "syllabic boundaries in fluent speech may be a fiction."
frequently random with respect to words or morphemes" (p. 173). They challenged proponents of syllabic segmentation to account for the way in which the sentence "He's a repeated offender" is parsed syllabically in a way compatible with its morphological structure. A partial solution to this challenge may come from the signal itself and/or from top-down constraints. Recent results reported by Mills (1980) suggest that word boundary information is coded in the syllable. Mills showed that a CVC target sequence which corresponds to a mono- or disyllabic word than either in this word spliced out of bisyllabic and trisyllabic words (candle, candlelight) or in these words themselves. Thus the word boundary cues in syllables would improve the mapping of syllable on the morphemes and words. Furthermore, top-down constraints may also help with this mapping process.

The evidence for syllabic segmentation presented here clearly has its implications for lexical access. As was pointed out earlier, the subjects' detection response probably precedes lexical access and thus is based on the prelexical code (Foss & Blank, 1980). Accordingly, the postulated syllabic segments could well serve as accessing units. This claim is at odds with lexical access models based on phonetic units. For these models, the words palmier and palace are still potential word candidates (in the same cohort) for the subject having heard the sequence /pall/. In the syllabic hypothesis, these two words could be distinguished earlier because their syllabic structure furnishes more information than was usually assumed.

The results reported here provide evidence for syllabic segmentation of speech. However, more research is necessary to determine which acoustic cues are actually used in syllabic segmentation and also in lexical access.

REFERENCES
Rubin, P., Turvey, M. T., & van Gelder, P. Initial phonemes are detected faster in spoken words than in spoken nonwords. Perception and Psychophysics, 1976, 19(5), 394-398.
Swinney, D. A., & Prather, P. Phonemic identification in a phoneme monitoring experiment: The