Perception & Psychophysics
1976, Vol. 19(5), 394-398

Initial phonemes are detected faster in spoken words than in spoken nonwords

PHILIP RUBIN and M. T. TURVEY
University of Connecticut, Storrs, Connecticut 06268
and
Haskins Laboratories, New Haven, Connecticut 06510

and

PETER VAN GELDER
University of Connecticut Health Center, Storrs, Connecticut 06268

Perception & Psychophysics
1976, Vol. 19(5), 394-398

Initial phonemes are detected faster in spoken words than in spoken nonwords

PHILIP RUBIN and M. T. TURVEY
University of Connecticut, Storrs, Connecticut 06268
and
Haskins Laboratories, New Haven, Connecticut 06510

and

PETER VAN GELDER
University of Connecticut Health Center, Storrs, Connecticut 06268

In two experiments, subjects monitored sequences of spoken consonant-vowel-consonant words and nonwords for a specified initial phoneme. In Experiment I, the target-carrying monosyllables were embedded in sequences in which the monosyllables were all words or all nonwords. The possible contextual bias of Experiment I was minimized in Experiment II through a random mixing of target-carrying words and nonwords with foil words and nonwords. Target-carrying words were distinguished in both experiments from target-carrying nonwords only in the final consonant, e.g., /bit/ vs. /bip/. In both experiments, subjects detected the specified consonant /b/ significantly faster when it began a word than when it began a nonword. One interpretation of this result is that in speech perception lexical information is accessed before phonological information. This interpretation was questioned and preference was given to the view that the result reflected processes subsequent to perception: words become available to awareness faster than nonwords and therefore provide a basis for differential responding that much sooner.

It is commonplace to conceptualize the process of pattern identification as a hierarchically organized sequence of operations that maps the structured energy at the receptors onto increasingly more abstract representations. In its most simplistic form, this conception characterizes the "conversation" between representations as unidirectional; that is, a more abstract representation is constructed with reference to a less abstract representation, but not vice versa. There are, however, a number of curious results that question the integrity of this characterization. By way of example, a briefly exposed and masked letter is recognized more accurately when part of a word than when part of a nonword (Wheeler, 1970; Reicher, Note 1). Other, related results suggest that this is a fairly general phenomenon. Thus, detection of an oriented line is significantly better when it is part of a briefly exposed, and masked, unitary picture of a well-formed three-dimensional object than when it is a part of a picture portraying a less well-formed, and flat, arrangement of lines (Weisstein & Harris, 1974). As revealed in the work of Biederman and his colleagues (Biederman, 1972; Biederman, Glass, & Stacy, 1973), this facilitation of "feature" detection by object context is matched by a facilitation of "object" detection by scene context: an object is more accurately and rapidly identified when part of a briefly exposed real-world scene than when it is part of a jumbled version of that scene, exposed equally briefly.

The present paper reports two experiments that were conducted to determine whether speech perception manifests phenomena analogous to those just described. In several recent experiments, a latency-of-detection task has been used to explore characteristics of speech processing. A case in point is the research of Foss and Swinney (1973), that demonstrated that two-syllable word targets were detected faster than their one-syllable counterparts, and that these in turn were detected faster than individual phonemes. Observations of this kind have motivated legitimate reservations about the relevance of the detection task to the analysis of perceptual stages. We will echo these reservations in our discussion. For the present, however, we draw attention to an important difference between the visual experiments described above and the speech experiments typified by Foss and Swinney (1973) (cf. McNeill & Lindig, 1973; Savin & Bever, 1970). The speech experiments have looked at differences in detection latencies for *different* kinds of targets (for example, phoneme, syllable). By way of contrast, the visual experiments

This research was made possible in part by support from NICHD Grant HD-01994. Reprints may be obtained from P. Rubin, Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510.

Table 1
Experiment I: Sample Test Sequences

JUT	LEG	*SIN*	RUG	WELL	RUN
MEG	GEEL	NUCK	HAEN	*BAL*	HIG
KEEJ	NUG	LAN	NAEN	*SIM*	DAJ
COME	*BAT*	LAG	TELL	TIN	GUM

Note—Target items are italicized.

have held the target type constant and varied the structure in which it is embedded. The question of interest has been the effect different structures have on the detection of their constituent elements. It is this question that provides the point of departure for our experiments.

There are several intimations that the global structure of a speech event significantly influences one's identification or detection of lower-order aspects of the speech signal. For example, it has been reported (Hadding-Koch & Studdert-Kennedy, 1964; Studdert-Kennedy & Hadding, 1973) that judgments about the final movements of a pitch contour—precisely, whether it rises or falls—are that it rises if the total contour is perceived as a question (even if the contour in fact has a final fall), and that it falls if the total contour is perceived as a statement (even if the contour in fact has a final rise). The present experiments examine phoneme targeting in words and nonwords, focusing on the initial phoneme. A demonstration that the detection of an *initial* phoneme of an utterance is affected by the lexical/semantic value of the utterance would dramatize the influence of holistic and higher-order properties on the detection of speech components.

EXPERIMENT I

Method

Subjects. The subjects were six male and nine female undergraduates at the University of Connecticut. The subjects participated to receive experimental credit in introductory psychology.

Materials and apparatus. Each subject received three blocks of trials, each block consisting of 16 sequences of consonant syllables. Sequences contained either six monosyllabic words or six monosyllabic nonwords, which conformed to the rules of English phonology. Within a block, there appeared an equal number of word and nonword sequences. Each sequence contained exactly one target syllable beginning with the phoneme /b/ or the phoneme /s/, occurring at a point between the second and the fifth syllable, with equal probability for each position in the sequence. Within a block there was an equal number of words and nonwords beginning with /b/ and /s/, and across there appeared an equal number of words and nonwords for each position in a sequence. The distinction between lexicon membership and nonmembership—that is, the word/nonword difference—was based on a change in the final consonant. No syllable, target or nontarget, contained a /b/ or /s/ in any position other than the initial position. Phonemes that are highly confusable with the target phonemes (e.g., /f/, /v/, /p/) did not appear in any syllable, target or nontarget. Target syllables were constructed so that the target was followed by different vowel phonemes; and to control for pronounceability, each of the different vowels followed the /b/ and /s/ targets an equal number of times. Table 1 presents some sample sequences. Items in this table are not given their phonetic spellings, but are presented in a form that makes clear the difference between words and nonwords. Presentation by block was either to the left ear, right ear, or both ears. A subject was presented with one of three ordered block sequences: (1) left, right, binaural; (2) right, binaural, left; (3) binaural, left, right. Presentation of particular targets in a sequence was also counterbalanced.

The speech was recorded at a normal speaking rate by a male speaker on one channel of an Ampex tape recorder. The speech waveform was then digitized and edited, using the Haskins Laboratories pulse-code modulation (PCM) system (Cooper & Mattingly, 1969). In the case of syllables beginning with the phoneme /s/, onset was standardized by starting sampling 100 msec before the start of the vowel. The recording of test tapes involved converting the digitized waveform samples to analog form. Using a Crown 800 tape recorder, test stimuli were recorded on one track of a test tape. The average duration of stimuli, both words and nonwords, was 450 msec. On the second track, a 500-msec, 500-Hz tone appeared coincident with the onset of target items. Syllables were separated by a 1-sec interstimulus interval and sequences of syllables were separated by 5 sec. The 500-Hz tone was used to start a timer in a Data General Nova computer which involved the computer sampling the signal from Track 2 of the presentation tape through an analog-to-digital converter. When a target item (and its coincident tone) was presented, the real-time clock in the computer was started and the button-push of a subject stopped the clock, thus giving the reaction time of the subject. Reaction times were printed out on an ASR-33 Teletype after each block. Reaction times greater than 1,500 msec were considered to be errors. The ear of presentation was controlled by feeding the Channel 1 output of a Sony TC-100 tape recorder through a mixer that put the signal in the left, right, or both speakers of two sets of Superex headphones—one set for the subject and one set for the experimenter to monitor the experiment.

Procedure. The subjects were told that they were going to hear sequences of monosyllables—both nonsense words and real words. Examples were then given both of syllables and sequences of syllables. The subjects were instructed to press a key as fast as possible whenever they heard a word or nonsense word that began with the sound /b/ or /s/. Further examples were given. The subjects were also instructed that their performance was being monitored, and that they should continue targeting even if they made an error. They then heard a practice block of eight sequences to familiarize them with the task. Five seconds before each block, the word "ready" (recorded on Track 1 of the test tape) was presented, to prepare the subjects for the beginning of the block. Before the start of each block, the subjects were informed of the ear of presentation and were again cautioned to wait for the "ready" signal and encouraged to push the key as fast as possible on hearing the target sounds. Between each block, the subjects were given a 2-min rest period.

Results

The mean reaction times across subjects for all conditions are presented in Table 2. A repeated measure analysis of variance was performed on the data. The only main effect to attain statistical significance was that of initial phoneme. The subjects responded significantly faster if the item, word or nonword, began with the phoneme /b/ ($\bar{X}$ = 600 msec) than if the item began with the phoneme /s/ ($\bar{X}$ = 736 msec) [$F(1,14) = 16$, $p < .01$]. The word vs. nonword difference was not significant,

although there was a tendency for subjects to respond faster to words ($\bar{X}$ = 683 msec) than to nonwords ($\bar{X}$ = 713 msec) [$F(1,14) = 3.33$, $p < .06$]. There was also a marked, though nonsignificant, tendency for binaural listening to be superior to monaural listening [$F(2,28) = 3.3$, $p < .1$]. The overall error rate was 8.3%, but an analysis of variance of the error data revealed no significant differences between the groups. Out of 720 total responses, there were 30 errors in both the word and nonword conditions. There was, however, a greater overall percentage of errors in items beginning with /s/ (9.72%) than in those beginning with /b/ (6.94%).

The greater difficulty in targeting for /s/ than for /b/ in initial position of a spoken item has been reported previously (Savin & Bever, 1970). A possible source of this difficulty in the present experiment was a reduction of sound quality in the /s/ segments. This was due to the sampling rate of the analog-to-digital converters in the Haskins PCM program.[1] Because of these factors, an analysis of variance independent of /s/ target data was undertaken. This analysis, for only those items with the phoneme /b/ in initial position (see Table 2), revealed that subjects responded significantly faster for words ($\bar{X}$ = 642 msec) than for nonwords ($\bar{X}$ = 678 msec) [$F(1,14) = 4.88$, $p < .05$]. Effects of ear of presentation and the interaction of ear of presentation and word vs. nonword did not attain significance.

EXPERIMENT II

Although the word/nonword effect was only marginally significant, the results of Experiment I support the notion that the higher order properties of a speech event affect the detection of its constituent parts. A second study was designed to further examine this effect using a slightly altered experimental procedure. The change in design represented, in part, an attempt to eliminate the dependence of phoneme targeting on the sequencing of meaningful or nonsense items within a block, that is, on extraneous contextual considerations. Further, the experiment was designed to increase the data base, while also randomizing the predictability of appearance of stimulus items. The last change involved the use of syllables with /s/ in the initial position as foils instead of as target items. This change resulted from the difficulty of determining exact onset of the phoneme /s/ and from the technical difficulty discussed above.

Table 2
Experiment I: Mean of Subjects' Mean Reaction Times in Milliseconds

	Initial Phoneme /b/ Ear of Presentation			Initial Phoneme /s/ Ear of Presentation		
	Left	Right	Binaural	Left	Right	Binaural
Word	648	669	608	744	716	714
Nonword	692	675	668	768	769	704

Table 3
Experiment II: Mean of Subjects' Mean Reaction Times in Milliseconds for /b/ in Initial Position

	Ear of Presentation		
	Left	Right	Binaural
Word	589	609	580
Nonword	645	655	631

Method

Subjects. The subjects were 19 female and 11 male undergraduates at the University of Connecticut. Subjects participated to receive experimental credit in introductory psychology.

Materials and apparatus. The stimuli in this experiment differed from those in Experiment I only in terms of the organization of a block. Each block consisted of 60 items. Within a block, there was an equal number of target items with the phoneme /b/ in initial position, foil items with /s/ in initial position, and foil items with various other consonants in initial position. These three types of items were equally divided into words and nonwords. In any word/nonword pair, the form of distinction consisted solely of a change in the final consonant (e.g., /bit/ vs. /bip/). All items were randomly organized throughout a block, and each block contained a different order of items. Interstimulus intervals were randomly assigned durations of 1, 2, 3, 4, or 5 sec. The practice block contained 18 items drawn from the overall stimulus set, organized analogously to an actual test block. The methods of stimulus presentation and of data collection were the same as in the previous experiment.

Procedure. The procedure in Experiment II differed from that of Experiment I in only two ways. First, subjects were informed of the nature of the block organization and the stimulus materials. Second, subjects were instructed to press, as fast as possible, one of two keys when they heard any syllable that started with the phoneme /s/ and to press the other key when they heard any syllable that began with the phoneme /b/. The subjects were required to target for syllables beginning with /s/ in addition to those beginning with /b/ in order to circumvent the possibility of rapid pressing for any utterance. However, as already noted, responses for /s/ items were not considered for analysis for the reasons cited above. The keypress procedure consisted of keeping the index finger on a start marker and moving upward and left or upward and right to the appropriate keys. The relation between phoneme and key was counterbalanced across subjects.

Results

Table 3 contains a summary of mean reaction times across subjects for all conditions. In a repeated measure analysis of variance, the only main effect to attain significance was that of word vs. nonword [$F(1,29) = 69.00$, $p < .001$]. Subjects responded faster to targets in words ($\bar{X}$ = 593 msec) than to targets in nonwords ($\bar{X}$ = 644 msec). A hint of a similar effect in the detection of final consonants is provided by the work of Steinheiser and Burrows (1973). Once again, there was a tendency for subjects to respond more quickly in the case of binaural

presentation. The overall error rate was 2.4%. Out of a total of 1,800 responses, 22 errors occurred in the nonword condition and 21 in the word condition.

DISCUSSION

As remarked at the outset, perception can be characterized as an orderly progression of mappings from less to more abstract representations. In the case of speech, these representations may be identified as follows: auditory, phonetic, phonological, lexical, syntactic, and semantic (Studdert-Kennedy, Note 2). This hierarchy suggests that a response to a particular phoneme could be initiated by the results of processing at the second representational level. But our experiments have shown that the lexical or semantic value of the speech utterance—that is, whether it is a word—affects the latency of phoneme detection. Should we take this to mean that the processing levels we have described are incorrectly arranged? That, contrary to the foregoing account, the lexical and semantic representations, say, precede the phonetic in the temporal course of speech perception, and that the phoneme, therefore, is not a major perceptual unit? Though these conclusions seem anomalous to many students of perception, there are some who would not be especially upset. Both Gibson (1966) and Kolers (1972; Kolers & Perkins, 1975), for example, abhor (for radically different reasons) accounts of perception couched in the language of atomistic elements and rules. For them, the search for, and arguments about, perceptual units are misguided, as is the treatment of perception as discrete and steplike. In their view, perhaps, our result and the conclusions it suggests about speech perception are not so much anomalous as they are indicative of the inadequacy of the hierarchic, elementaristic formulation of perception.

The necessity for including the above caveat in the present discussion depends in part on the assumption that our experiments actually tap the perception of speech. There is a possibility that this assumption is false; in short, that our experiments do not comment on the nature of perceptual processing at all.

In response to the data obtained from latency-of-detection experiments, Foss and Swinney (1973) drew the following, speculative conclusion: the processes of perceiving are largely separate from and independent of the processes by which perceived events become consciously identified to serve as a basis for differential responding. One of us (Turvey, 1974) has drawn a similar distinction, but in Polanyi's (1964, 1966) terms, that is, between the processes of tacit and explicit knowing. Turvey (1974) conjectures that the processes underlying tacit knowing (perceiving) are different in kind from those underlying explicit identification. From this point of view, the detection task does not reveal perceptual units, nor does it tap perceptual processes; to the contrary, it assays operations subsequent to perception. These operations of identification can result in differences in the rates at which descriptions of linguistic events become available to consciousness as a basis for responding (cf. Ball, Wood, & Smith, 1975; Foss & Swinney, 1973).

Consider, once again, the perception of speech from a hierarchical point of view. There are a number of reasons for proposing that the relations among the levels ought to be quite flexible, with higher-level procedures correcting or verifying descriptions reached tentatively by lower-level procedures (cf. Studdert-Kennedy, Note 2). Given the outcome of our experiments, therefore, we can assume from this perspective that there has been considerable confluence among the various levels prior to that identification of the speech event permitting differential responding. But assuming that the representations are at least partially successive and that phoneme detection occurs early on, then why should the response contingent on phoneme detection be delayed until both lower and higher representations have made their contribution? The answer, apparently, is that in the detection task the response mechanism cannot be engaged until perception is complete. In short, even if there do happen to be levels or stages in speech perception, the detection task, unfortunately, will not reveal their order of operation to us. Let us return, therefore, to Foss and Swinney's (1973) thesis.

A cardinal feature of their argument, and one that bears on our particular finding, is that larger language units become available to consciousness, that is, become explicit, sooner than smaller language units (cf. Ball et al., 1975). In the present experiments, all items, words, and nonwords, were monosyllables. If we take the notion of "larger units" literally, then we cannot attribute the latency difference in initial phoneme detection to a hypothesized word/nonword difference in time to access consciousness, since words and nonwords were of the same size. Consequently, if words and nonwords become available to consciousness at the same latency, then the word advantage effect in initial detection must be due to a difference in the ease with which words and nonwords as "larger units" can be fractionated into phonemes as "smaller units." However, an alternative interpretation, and one that we prefer, is that familiarity and/or meaning, rather than size, determines the latency of availability to consciousness. In this view, because a word has meaning, its description is made

available sooner than that of a nonword (cf. Avant & Lyman, 1975). The latency difference for initial phoneme detection can then be attributed to this differential rate of availability to awareness of the linguistic description.

REFERENCE NOTES

1. Reicher, G. M. *Perceptual recognition as a function of meaningfulness of stimulus material.* Technical Report No. 7, 1968. The University of Michigan, Human Performance Center.

2. Studdert-Kennedy, M. *Speech perception.* Haskins Laboratories Status Report on Speech Research SR-39/40, 1974, 1-52.

REFERENCES

Avant, L. L. & Lyman, P. J. Stimulus familiarity modifies perceived duration in prerecognition visual processing. *Journal of Experimental Psychology: Human Perception and Performance*, 1975, 1, 205-213.

Ball, F., Wood, C., & Smith, E. E. When are semantic targets detected faster than visual or acoustic ones? *Perception & Psychophysics*, 1975, 7, 1-8.

Biederman, I. Perceiving real-world scenes. *Science*, 1972, 177, 77-79.

Biederman, I., Glass, A. L., & Stacy, E. W., Jr. Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 1973, 97, 22-27.

Cooper, F. S., & Mattingly, I. G. A computer controlled PCM system for the investigation of dichotic perception. *Journal of the Acoustical Society of America*, 1969, 46, 115(A).

Foss, D. J., & Swinney, D. A. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, 1973, 12, 246-257.

Gibson, J. J., *The senses considered as a perceptual system.* Boston: Houghton Mifflin, 1966.

Hadding-Koch, K., & Studdert-Kennedy, M. An experimental study of some intonation contours. *Phonetica*, 1964, 11, 175-185.

Kolers, P. A. *Aspects of motion perception.* New York: Pergamon Press, 1972.

Kolers, P. A., & Perkins, D. N. Spatial and ordinal components of form perception and literacy. *Cognitive Psychology*, 1975, 7, 228-267.

McNeill, D., & Lindig, K. The perceptual reality of phonemes, syllables, words and sentences. *Journal of Verbal Learning Verbal Behavior*, 1973, 12, 419-430.

Polanyi, M. *Personal knowledge: Towards a post-critical philosophy.* New York: Harper, 1964.

Polanyi, M. *The tacit dimension.* Garden City, N.Y: Doubleday, 1966.

Savin, H. B., & Bever, T. G. The nonperceptual reality of the phoneme. *Journal of Verbal Learning and Verbal Behavior*, 1970, 9, 295-302.

Steinheiser, F. H., Jr., & Burrows, D. J., Chronometric analysis of speech perception. *Perception & Psychophysics*, 1973, 13, 426-430.

Studdert-Kennedy, M., & Hadding, K. Auditory and linguistic processes in the perception of intonation contours. *Language and Speech*, 1973, 16, 293-313.

Turvey, M. T. Constructive theory, perceptual systems, and tacit knowledge. In W. B. Weimer & D. S. Palermo (Eds.), *Cognition and the symbolic processes.* Hillsdale, N.J: Lawrence Erlbaum Associates, 1974.

Weisstein, N., & Harris, C. S. Visual detection of line segments: An object-superiority effect. *Science*, 1974, 186, 752-754.

Wheeler, D. D. Processes in word recognition. *Cognitive Psychology*, 1970, 1, 59-85.

NOTE

1. The sampling rate used in the Haskins PCM system was 8 kHz when the experiment was conducted, yielding a maximum effective bandwidth of 4 kHz. In addition to this Nyquist limit, a further restriction was imposed by the bandpass filtering, which made the true effective range of the signal 90 Hz to 3.2 kHz. This range does not allow for an entirely adequate representation of fricatives such as /s/. The most recent Haskins PCM system can sample the signal at the rate of 20 kHz, with bandpass filtering making the effective range 90 Hz to 8.4 kHz.

(Received for publication November 7, 1975;
revision accepted February 8, 1976.)