Chronometric analysis of speech perception*

F. H. STEINHEISER, JR.[†] and DAVID J. BURROWS^{††}

Center for Research in Human Learning, University of Minnesota, Minneapolis, Minnesota 55455

The relationship between the phonological properties of speech sounds and the corresponding semantic entries was studied in two experiments using response time measures. Monosyllabic words and nonsense words were used in both experiments. In Experiment I, Ss were each presented with individual items and were required, in three different conditions, to respond positively if (1) the item contained a particular final consonant, (2) the item was a real word, (3) the item contained either a particular consonant or was a real word. Latencies indicated that separate retrieval of phonological and lexical information took about the same time, but that their combined retrieval was longer, indicating a serial or overlapping process. In Experiment II, Ss were presented with pairs of items, and they responded positively if (1) the two items were physically identical, (2) the two items were lexically identical (both real words or both nonsense words). Response latencies were longer for lexical than for physical matches. Lexical matches were significantly slower than physical matches even on the same pair of items. The results imply differential accessibility to separate loci of phonological and semantic features.

An information-processing approach to perception assumes that various stages of decoding are required to interpret a stimulus (Posner & Mitchell, 1967; Haber, 1969). It is often possible to measure these stages by obtaining response latencies (Sternberg, 1969). For example, a S could be presented with a spoken item such as *get*, and asked to indicate whether it ended in a /t/ (as opposed to alternatives like /m/ or /b/). In another task, the S could indicate if the item *get* was a real word (as opposed to a nonsense word like *geb*). If latencies in the first task were longer, we might infer that the item was first perceived in its entirety as a syllable or word and then analyzed into its constituent phonemes. If the naming task produced longer latencies, we might then infer that this additional time reflected a semantic search in which a lexical entry was or was not located for that speech sound. But if there were no differences between the two tasks, it could be argued that the syllable was the encoding unit for speech sounds, from which phonological features are decoded in the same length of time as are semantic features. A disjunctive task, in which either a target phoneme or a real word was the target, was used to resolve this ambiguity in the first experiment.

Lieberman (1970) suggests that speech sounds are not simple strings of auditory stimuli or phonemes, but rather comprise a complex code. For this reason, the argument has been made that units larger than the

phoneme, most likely syllables, could be the fundamental unit in decoding a speech signal: "To find acoustic segments that are in any reasonably simple sense invariant with linguistic (and perceptual) segments—that is, to perceive without decoding—one must go to the syllable level or higher [Lieberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, p. 451]."

Phonemes, as minimally distinguishing units that denote "nothing but mere otherness," are essential for perceiving speech. For most Os, the difference between two different phonemes is greater than that for the "same" phoneme in different contexts. That is, the perceived difference between /d/ and /b/ (as in *dig* and *big*) is greater than that for the /b/s in *big* and *bag*. But even if phonemes should be identified subsequent to syllables, some kind of phonemic analysis must logically precede word recognition. For example, in a minimally different pair, like *bat* and *bad*, one must discriminate the phonemes prior to identifying the words.

The present experiments attempted to decompose the process of perceiving the meaning of words into substages. It would intuitively appear that this should be a serially ordered process which starts with the identification of constituent phonemes, which are then combined into syllables. The syllables are then combined into words, with the meaning of each word being obtained by some sort of lexical look-up. Recent evidence, however, brings into question the assumption that speech processing starts with the identification of individual phonemes. In one study, Savin and Bever (1970) presented Ss with a set of nonsense monosyllables. The S responded as soon as he heard a designated target in the taped list. Response times were significantly faster to a target that was a complete syllable than to a target that was a phoneme in that syllable. Warren (1971) presented his Ss with passages of connected discourse and random strings of words.

*This experiment was supported by NIMH postdoctoral fellowships to the authors and by grants to the University of Minnesota, Center for Research in Human Learning, from NSF (GB-17590), from NICHD (HD-01136 and HD-00098), and from the Graduate School of the University of Minnesota.

[†]Present address: The J. F. Kennedy Habilitation Institute, The Johns Hopkins University, 707 N. Broadway, Baltimore, Maryland 21205.

^{††}Present address: Psychology Department, State University of New York, Brockport, New York 14420.

Identification times for monosyllabic words were consistently faster than identification times for constituent phoneme clusters. Furthermore, latencies to phoneme clusters were shorter than latencies to individual phonemes within the clusters. (For example, the /to/ in *stone* was identified more rapidly than the single phoneme /t/.)

Furthermore, there is some evidence that semantic interpretation and phoneme identification interact in a complex fashion rather than being serially ordered independent processes (Foss, 1969). Ss were run in a phoneme monitoring task in which the target phoneme /b/ was to be identified as rapidly as possible. The target followed either low-frequency or high-frequency synonymous words in sentences: "... itinerant bassoon player" vs "... travelling bassoon player." Latencies were longer when identifying the target after low-frequency words.

The first experiment attempted to test the hypothesis that phoneme identification must be prior to (and therefore faster than) word identification. A phoneme monitoring technique was used, similar to that employed by Foss, except that the *terminal* phoneme of an isolated monosyllable was the target. The set of monosyllables consisted of real words and nonsense words. In another condition, the task was to indicate whether the item was a word, regardless of the terminal phoneme. Presumably an additional processing stage would be required to extract meaning (semantic features) after the constituent phonemes had been identified. Hence, it should take longer to identify that, e.g., *get* is a real word than to indicate in the terminal phoneme monitoring condition that *get* ends in a /t/—the target phoneme.

The second experiment employed a successive matching procedure (Posner & Mitchell, 1967), in which conceptual identity and physical identity were the criteria for an affirmative match. For example, in the physical identity condition of Experiment II, presentation of *get*, *get* would require a positive response, but *get*, *guess* would require a negative response. In the semantic identity condition, presentation of *get*, *get*, or *get*, *guess* would require a positive response, since meaningful words comprise each pair, although *get* and *guess* are not identical words. If physical comparisons require fewer comparison stages than semantic (conceptual) comparisons, then response times to the physically same pair of items should be shorter in the physical identity than in the semantic identity condition.

EXPERIMENT I

Method

Subjects

Ten females and 14 males, selected from psychology courses at the University of Minnesota, participated in the experiment

either for extra credit points in the course or for \$1.80 per session. Two additional Ss were dropped because of poor performance.

Apparatus

Tape recordings of various monosyllabic words and nonsense words were made by E on a Sony TC-355 tape deck at a tape speed of 19 cm/sec. S wore a pair of Koss Pro-4A stereo headphones. A Beckman-Berkeley timer started automatically at the onset of each taped item, within a range of 5 msec. Depression of either of two response keys by S stopped the timer, with the elapsed time being printed on a paper strip. An event recorder showed which button was pushed on each trial.

Stimuli

Three tape recordings were made, one for each condition. Tape 1 consisted of 16 real monosyllabic words which ended in the consonant /t/ (e.g., *late*); 16 nonsense monosyllables which were derived from the corresponding real words by simply changing the vowel (e.g., *loat*); 16 real words which had the same initial consonant and vowel as the first set of words but did not end in /t/ (e.g., *lace*); and 16 nonsense words which had the same initial consonant and vowel as the first set of corresponding nonsense words but did not end in /t/ (e.g., *loce*). This tape was used only in the phoneme monitoring condition.

Tape 2 consisted of the same set of real words as was used in Tape 1, with the first set of nonsense items now having the same initial consonant and vowel as the real words, but ending in a voiced stop consonant (e.g., *labe*), and the second set of nonsense items ending in a voiced fricative (e.g., *lave*). This tape was used only in the word monitoring condition. (Thus, S had to attend to the entire item, including the final consonant, before deciding if it was a word.)

Tape 3 consisted of the same set of items as Tape 1, except that the nonsense items that did not end in /t/ were repeated three times. This was done to achieve equiprobability of response in this condition.

Tapes 1 and 2 consisted of 64 items, and Tape 3 consisted of 96 items. The ordering of items within a tape was random. A 1,000-Hz warning tone of .5 sec duration preceded each item by 1.5 sec. The time interval between successive items was approximately 10 sec.

Procedure

Each of the 24 Ss was tested individually in three conditions on separate days, counterbalanced for order across Ss. In the phoneme monitoring condition, S listened to Tape 1, and was instructed to push one button if the item ended in /t/ and to push another button if the item ended in a different consonant. In the word monitoring condition, S listened to Tape 2, and was asked to push one button if the item was a real word and to push another button if the item was a nonsense word. In the disjunctive condition, S listened to Tape 3, and was asked to push one button if either a real word, regardless of its ending, or a nonsense word that ended in /t/ was presented and to push the other button if the item was nonsense word which did not end in /t/.

At the start of each session, E reviewed with S the list of items that would be used in that session. After listening to E's pronunciation, S was asked to pronounce each item on the list. E corrected S when necessary. After S was able to pronounce the list of nonsense items reliably, 20 practice trials were taken. S was allowed to adjust the volume control during the practice trials. After practice, S was shown his record of latencies and errors.

The session was run in two parts, separated by a 5-min rest break. The same button position was maintained for each S throughout the three sessions, with the "positive" button being

Table 1
Mean RTs (Milliseconds) and Standard Deviations for Each Condition Over All Ss

Condition	Type of Item							
	Word /t/		Nonword /t/ (Non /t/)		Word Non /t/		Nonword Non /t/	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Phoneme	761	12	792	14	805	10	835	10
Word	787	13	<i>814</i>	<i>16</i>	814	20	860	15
Disjunctive	771	13	848	13	910	11	925	18

Note—*Italicized entries in the word condition indicate that nonwords not ending in /t/ were used.*

on the left for 12 Ss and on the right for the other 12 Ss. The experimental design was a randomized block factorial (Kirk, 1969), with three levels of instruction and four levels of items. Assignment of Ss to an order of the instructions was counterbalanced.

Results

The error rates for the phoneme, word, and disjunctive conditions were 2.8%, 3.3%, and 4.9%, respectively. Trials on which an error occurred were omitted from further analysis.

To reduce the positive skewness typically found in response latency data, the longest time to each of the four item types was deleted from each S's data. In the disjunctive condition, the three slowest times were deleted for nonsense words which did not end in /t/.

The average latency across all 24 Ss for each of the four stimulus groups in each of the three conditions was then computed. Table 1 shows each mean and its standard deviation. Each entry is based upon approximately 340 responses, except for the non-/t/ items in the disjunctive condition, 925 msec, based upon approximately 1,000 responses.

A repeated measures analysis of variance revealed a significant effect due to the type of instructions, $F(2,46) = 12.18$, $p < .001$, type of item, $F(3,69) = 85.84$, $p < .001$, and Instruction by Item interaction, $F(6,138) = 10.89$, $p < .001$. All tests on simple main effects were significant at $p < .001$, except the two tests for real words and nonsense words which ended in /t/. These latter two were not statistically significant.

Inspection of Table 1 shows that the average response times were fastest under all three sets of instructions to real words which ended in /t/ and slowest to nonwords which did not end in /t/. Subsequent interviews with Ss confirmed that each S had difficulty with at least one nonsense item because it "sounded strange" or because it was confusable for that S with a real word. Items which ended in /t/ were said to be more "distinctive" than items which ended in other consonants, such as /s/ and /f/.

Further inspection of Table 1 shows that there was an average difference of only about 20 msec between latencies to the four item types in the phoneme monitoring condition and the four item types in the word monitoring condition. A Tukey HSD test showed

that this difference was not statistically significant, $q(3,69) = 2.1$, $p < .05$. The difference between the disjunctive condition and the average of the phoneme and word monitoring conditions is 49 msec. A Scheffé test showed that this difference was marginally statistically significant, $F(2,46) = 10.98$, $p < .05$.

There appears to be a trend in the disjunctive condition, in that it took about 75 msec longer to identify a terminal /t/ in nonwords than in real words, and an additional 62 msec to identify a real word as real when it did not end in /t/. Tukey HSD tests yielded $q = 3.25$, $p < .05$, and $q = 3.93$, $p < .01$, respectively.

Discussion

These results suggest that identification of a terminal phoneme from a monosyllable is no faster than a decision as to whether the monosyllable is a word. However, the time to accomplish a disjunctive task is much greater. This implies that the two separate tasks are not dependent on each other. If phoneme monitoring were a component of word monitoring, then the time to accomplish word monitoring would be greater than the time for phoneme monitoring. The greater times in the disjunctive task further suggest that the two tasks are separate: doing both tasks takes longer because accomplishing one task does not "save" any of the time required to accomplish the other task. Thus interpreted, the data agree with the results of Savin and Bever (1970) and Warren (1971), that the phoneme is not the *primary* unit of speech analysis.

Furthermore, it appears that lexical information may be accessed simultaneously during the process of phoneme identification. One likely possibility is that processing begins with some more complex unit, such as a syllable, which can then be analyzed semantically, to find a corresponding lexical entry, or phonologically, as it is decomposed into constituent phonemes. Phoneme identification and word recognition would then be separate processes, yielding the obtained results.

EXPERIMENT II

Experiment II was an attempt to test the hypothesis that the first stage of processing involves perception at the syllabic level, with lexical information being accessed

pairs in the lexical identity condition was 923 msec, and in the physical identity condition, 710 msec. This difference, 213 msec, was statistically significant, based on a Scheffé test: $F(1,38) = 43.1$, $p < .001$.

Response times were consistently faster in the physical identity than in the lexical identity condition, for each of the six different types of item pairs. Each difference was statistically significant at the $p \leq .01$ level or beyond, based on Tukey HSD tests. For example, the difference in average response times to physically identical words between the two conditions is 140 msec ($826 - 686$). When evaluated by a Tukey HSD test, $q = 5.18$, which surpasses the .01 rejection level of $q' = 3.7$.

In the lexical identity condition, the average response time to a pair of identical real words and to a pair of identical nonwords was 868 msec $[(826 + 911)/2 = 868]$. The average response time to a pair of different real words and to a pair of different nonwords was 977 msec $[(956 + 1,001)/2 = 977]$. This difference, $977 - 868 = 109$ msec, was statistically significant by a Scheffé test, $F(1,38) = 11.2$, $p < .01$. This time of 109 msec most likely represents a pure "lexical look-up" state within the lexical identity condition after a physical matching stage has failed to yield a positive match.

The results of Experiment II suggest that the processing of lexical and phonological information take the same amount of time. Furthermore, the accessing of lexical information is slower than syllabic encoding. These conclusions are consistent with the interpretation that syllables are directly perceived, with "meaning" being accessed on the basis of a look-up in the internal lexicon where the entire syllable is the address. Phonemes themselves seem to get bypassed in this process.

GENERAL DISCUSSION

The results of Experiment I suggest that processing of phonological information does not necessarily precede the processing of semantic information. If encoding proceeded from phoneme to syllable to word to meaning, there should have been a greater difference between the phoneme and word monitoring tasks. The results imply that some other unit is directly perceived and that this unit can then be analyzed one way to obtain phonemic information or another way to obtain semantic information. The longer times for the disjunctive condition support the hypothesis that phonemic and semantic information are processed separately, since this combined search task produced longer times than either simple task alone.

The results of Experiment II support the hypothesis that the basic perceptual unit may indeed be the

syllable, since syllabic information can be handled more rapidly than semantic information. This suggests that semantic information is organized so that it is addressed with syllabic information. The time for this addressing is reflected in the differences for the two tasks in Experiment II. Further, it suggests that syllables are directly perceived without being synthesized from "lower-order" units, such as phonemes. This is not to deny the importance of the concept of phonemes, since differences of one phoneme are critical in speech processing, e.g., *bat* and *bad* have different meanings. Rather, what is implied is that all of the information in a syllable is processed together, so that syllable perception cannot be broken down into serially ordered steps.

The hypothesis that the syllable is a more basic perceptual encoding unit is consistent with the results of similar experiments (Savin & Bever, 1970; Warren, 1971), and is also consistent with other speech perception research (Liberman et al, 1967; Liberman, 1970). Liberman and his colleagues at the Haskins Laboratories propose that there are no segmentable units in the acoustic stream which correspond to individual phonemes, while syllable segmentation is connected with physical aspects of the acoustic stream. The syllable may therefore be considered as a physically defined unit, and the phoneme as a truly abstract concept representing a stage not actually passed through in the processing of speech.

REFERENCES

- Foss, D. Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning & Verbal Behavior*, 1969, 8, 457-462.
- Haber, R. N. (Ed.) *Information-processing approaches to visual perception*. New York: Holt, Rinehart, & Winston, 1969.
- Kirk, R. *Experimental design: Procedures for the behavioral sciences*. Belmont, Calif: Brooks/Cole, 1969.
- Liberman, A. The grammars of speech and language. *Cognitive Psychology*, 1970, 1, 301-323.
- Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, 74, 431-461.
- Posner, M. I., & Mitchell, R. F. Chronometric analysis of classification. *Psychological Review*, 1967, 74, 392-409.
- Savin, H. B., & Bever, T. G. The nonperceptual reality of the phoneme. *Journal of Verbal Learning & Verbal Behavior*, 1970, 9, 295-302.
- Sternberg, S. The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica*, 1969, 30, 276-315.
- Warren, R. M. Identification times for phonemic components of graded complexity and for spelling of speech. *Perception & Psychophysics*, 1971, 9, 345-349.

(Received for publication September 5, 1972;
revision received January 15, 1973.)