JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 15, 73–83 (1976)

Units of Speech Perception: Phoneme and Syllable

ALICE F. HEALY
Yale University

AND

JAMES E. CUTTING
Wesleyan University and Haskins Laboratories

Two detection experiments were conducted with short lists of synthetic speech stimuli where phoneme targets were compared to syllable targets. Unlike previous experiments heterogeneous lists of syllables and phonemes were used to remove possible bias created by homogeneous lists. In Experiment I, targets that matched the response items in linguistic level were recognized faster than those that mismatched, whether the targets were syllables or phonemes. In Experiment II, where all targets and response items matched in level, phonemes were recognized faster than syllables when phonemes were relatively easy to identify, but the reverse held when phonemes were harder to identify. These results suggest that phonemes and syllables are equally basic to speech perception.

Phonemes and syllables constitute two levels in the linguistic hierarchy. One simple information processing model of speech perception (see for example McNeill & Lindig, 1973, for a discussion of this model) holds that the phonemic level is basic and that phonemes are the perceptual units for speech. According to such a model, we process speech by entering the language hierarchy at the bottom and working upward through the hierarchy (see Miller, 1962): We first recognize phonemes and then use this analysis to recognize higher-level linguistic entities such as syllables, words, and so on. Savin and Bever (1970) rejected this simple model, and any analogous model where phonemes are recognized before syllables but some lower-level features are recognized before phonemes, on the basis of an experiment where subjects responded faster to syllable targets than to phoneme targets when they searched through a list of syllables. They concluded that phonemes are recognized only after the recognition of syllables, or that the units of speech perception are at a level higher than the phoneme. In other words, Savin and Bever suggested that speech perceivers gain entry into the linguistic hierarchy at a level above phonemes, and that phonemes are only subsequently decomposed from higher-level units. However, McNeill and Lindig (1973) suggested that Savin and Bever's results were artifactual, stemming from their choice of list construction. Specifically, McNeill and Lindig showed that subjects respond faster to those targets that match the linguistic level of the items in the list, whatever that level. Subjects are faster at detecting syllables than phonemes in a list of syllables, but they are faster at detecting phonemes than syllables in a list of phonemes. One possible interpretation of these results is that the level in the linguistic hierarchy at which speech perceivers focus their attention depends on the level of the search list as well as

The research reported here was supported in part by Grant HD-01994 from NICHD to the Haskins Laboratories and PHS Grants MH26573 and RR0-7015 to Yale University. The authors are indebted to R. Crowder, W. K. Estes, A. Liberman, D. McNeill, and G. A. Miller for helpful discussions about this research. Requests for reprints should be sent to Alice F. Healy, Department of Psychology, Yale University, New Haven, Connecticut 06520.

the level of the target so that the speech perceiver must divide his attention between two levels whenever the target and search list do not match in level.

The present study reconsiders the question which level the speech perceiver enters the linguistic hierarchy by removing the matching factor which seems to have obscured results relevant to this question in the previous studies. This effect was achieved by employing heterogeneous rather than homogeneous lists of items. Specifically, subjects listen to lists that include a mixture of syllables and phonemes. Unlike previous studies, the nature of the list in the present study should not bias the subject to focus his attention at either level. In addition, the present study differs from the previous ones by using synthetic speech which allows finer control over the stimuli than was previously available.

Experiment I

Method

Subjects. Sixteen students of Yale University with no known hearing defects participated in this experiment either on a voluntary basis or for course credit. The subjects were divided into two groups of eight depending on their time of arrival for testing.

Apparatus. The speech stimuli were generated on the Haskins Laboratories' parallel resonance synthesizer, which permits precise control of the pitch contour and duration of the stimuli as well as the amplitude and frequency level of the three component formants. The speech stimuli were digitized and stored on disk file. Stimulus tapes were constructed using the Haskins' pulse code modulation system, which insured that all instances of a given stimulus sound were identical. The stimuli were reconverted into analog form and recorded on one track of a dual-track Ampex AG500 tape recorder. The sounds on this track of the tape were transmitted to the subject binaurally via a listening station through a pair of Telephonics earphones (Model TDH39). Gains on the tape recorder and listening station were adjusted so that stimuli were presented at approximately 80 db re 20 $\mu N/m^2$.

On the second track of the tape a 100-msec 1000-Hz tone was recorded, aligned with the onset of every item to which the listener was to respond. This tone triggered an electronic counter (Hewlett-Packard Model 522B) which stopped when the subject depressed a telegraph key. The elapsed time between tone onset and the subject's response was transcribed from a Hewlett-Packard Digital Recorder (Model 560A). The telegraph key was also connected to a small light, visible to the experimenter but not to the subject, which designated whether the telegraph key was depressed. In this manner, the experimenter detected false alarm responses.

The subject sat alone in a sound treated room, facing a table which supported the telegraph key, while the experimenter sat in an adjacent room which housed the timer and the other control devices.

Design and materials. Fifteen speech stimuli were generated, They consisted of five vowel stimuli /i, æ, I, aI, eI/, and 10 vowel-consonant stimuli /it, iv, æt, æn, In, Id, aIv, aIm, eId, eIm/ (see Bronstein, 1960). The vowel stimuli will hereafter be considered phonemes, whereas the vowel-consonant stimuli will be considered syllables. Unlike the study of McNeill and Lindig (1973), the present study used vowels instead of consonants as phonemes since vowels can be produced more easily in isolation.[1] Note that the 10 syllable stimuli can be divided into five pairs where each stimulus in the pair begins with the same vowel, or can be divided into five different pairs where each stimulus in the pair ends with the same consonant. All stimuli were identical in duration (325 msec), pitch contour (rising from 116 to 124 Hz in the first 40 msec, then falling linearly to 82 Hz by the end of the stimulus),

[1] McNeill and Lindig's phonemes were never produced in isolation; they were always part of consonant-vowel syllables where the vowel was held constant.

and overall amplitude. Monophthongal vowels /i, æ, I/ were steady-state throughout, whereas diphthongal vowels /aI, eI/ contained transitions from initial vowel nucleus to final vowel target (110 to 260 msec into the stimulus). Final consonants in the syllable stimuli occurred throughout the final 125 msec of those stimuli; the initial 200 msec of the syllable stimuli, however, were identical to the corresponding phoneme stimuli.

Four stimulus tapes were constructed for the two tasks of the present experiment—identification and detection. There was one tape for the identification task, and one practice tape and two experimental tapes for the detection task. The identification tape consisted of a series of 75 items with 3.0 sec between items, including five instances of each of the 15 stimuli. The first 15 items included each of the 15 different stimuli, and the following 60 items included a random sequence of four instances of each of the 15 stimuli.

Both experimental tapes for the detection task consisted of 80 trials. Each trial began with the statement *the target is*, followed by the target stimulus for the particular trial, the statement *here is the list*, and 2 sec later a four-item list. Intervals between items within the list were 2 sec each. The lists were composed of stimuli in a quasirandom sequence with several constraints: (*a*) Each list contained two syllables and two phonemes; (*b*) across the 80 trials each syllable occured 16 times and each phoneme 32 times; (*c*) four different vowels occurred in every list; (*d*) no consonant was repeated within a list; and (*e*) no consonant member of a target syllable appeared in any item in the list unless it appeared in the item to which the subject was supposed to respond.

The item in the list to which the subject was supposed to respond will henceforth be referred to as the "response item." The response item is identical to the target item when there is a *match* in linguistic levels, but the response item and target item differ when there is a *mismatch*. To use the terminology of McNeill and Lindig (1973) for mismatches, "upward search" occurs when the target is a phoneme and the response item a syllable, whereas "downward search" occurs when the target is a syllable and the response item a phoneme. The position of the response item within a list was quasirandom except that it never appeared in Position 1. It occurred 20 times in each of the Positions 2, 3, and 4. In the remaining 20 trials, there was no response item in the list.

The two detection tapes were identical except for the order of the target stimuli. Whenever a syllable stimulus occurred as the target on one tape, the corresponding phoneme stimulus (the phoneme stimulus with the same vowel as the syllable stimulus) occurred on the other tape. Within a tape, the order of target stimuli was quasirandom. Each of the five phoneme stimuli occurred eight times as targets and each of the 10 syllable stimuli occurred four times as targets. Each phoneme target occurred twice and each syllable target once for each of the three response item positions and for lists containing no response item. Thirty of the 60 trials with response items contained targets that matched the response items in linguistic level and 30 trials contained targets that mismatched the response items. However, the match-mismatch factor was not perfectly counterbalanced with respect to the variables of response item position, target level, and target vowel.

The practice tape was designed along the lines of the two experimental tapes used in the detection task except that only 10 rather than 80 trials were included.

Procedure. Each subject was run individually in an hour-long session. The session was divided into two parts: the identification task followed by the detection task. The two groups of subjects differed only in the experimental tapes employed during the detection task.

For the identification task, the subject was given two sheets of paper, one a numbered response sheet and the other a list of the 15 stimuli in the order they occurred initially on

the identification tape. The stimuli were named as follows: *eat* (/it/), *eve* (/iv/), *id* (/Id/), *in* (/In/), *an* (/æn/), *at* (/æt/), *aim* (/eIm/), *aid* (/eId/), *I'm* (/aIm/), *I've* (/aIv/), *ee* (/i/), *ih* (/I/), *aa* (/æ/), *A* (/eI/), *I* (/aI/). The subject listened to the first 15 stimuli in order to learn the sound-name correspondences, and then wrote down the name of each of the 60 subsequent stimuli on his response sheet.

The detection task included 10 practice trials followed by 80 experimental trials, with a short break at midtest. For each list the subject was told to press the response key as soon as he heard "the target sound itself, the initial sound in the target alone, or the target sound embedded in a longer sound." To help clarify these instructions, the subject was given the following example: "Imagine that the target is 'own'. You should respond either if you hear the sound 'own' itself or 'owe', the initial sound in 'own'. For a second example, imagine that the target is 'owe'. You should respond either if you hear the sound 'owe' or the longer sound 'own' which includes the sound 'owe'." The subject was further instructed that "on some lists the target sounds [response items] will not occur."

Instructions stressed the importance of both speed and accuracy of response. The subject was told that it was not necessary for him to wait to respond until he had heard the whole of a sound. Furthermore, as in the study by McNeill and Lindig (1973), the subject was told that his response should be to the pronunciation not the spelling of the sound. For clarification of this last point the subject was again referred to the example described above, "Given the target 'own', you should respond when you hear 'owe' even though 'owe' is included in 'own' according to pronunciation but not spelling." The subject was further instructed to make no more than one response during a given list.

Results and Discussion

Identification task. The results of the identification task are summarized in Table 1 in terms of error percentages pooled across subjects as a function of vowel type and stimulus level (phonemes vs. syllables). Two aspects of the data are noteworthy in light of the results of the detection task to be discussed below. First, the percentage of errors on the phoneme stimuli /I/ and /eI/ is significantly larger than on the phoneme stimuli /i/, /æ/, and /aI/ according to a Wilcoxon Matched-Pairs Signed-Ranks test, $T(7) = 0$, $p < .05$. Second, the error percentages on the phonemes are less than on the corresponding syllables for the vowels /i/, /æ/, and /aI/, whereas the reverse holds for /I/ and /eI/. These errors imply that the phoneme stimuli /I/ and /eI/, as they are actualized in the present experiment with synthetic speech, are more difficult to identify than the other three phoneme stimuli and are more difficult to identify than their corresponding syllable stimuli whereas the reverse holds for the other three vowels. This

TABLE 1

ERROR PERCENTAGES IN IDENTIFICATION TASKS OF EXPERIMENTS I AND II

Experiment	Stimulus level	Vowel /i/	/æ/	/aI/	/eI/	/I/
I	Phoneme	2	3	8	11	31
	Syllable	12	4	19	1	8
II	Phoneme	8	3	14	22	31
	Syllable	14	6	23	2	11

implication is interesting because these two phonemes are not those that are most complex acoustically. The diphthongs /aI/ and /eI/ which have moving formants are acoustically more complex than the other three monophthongal phonemes. Although /eI/ was shown to be relatively difficult to identify, /aI/ was not found to be especially difficult. Most likely the ease of identification of a particular phoneme depends upon its similarity to the other phonemes in the population rather than its absolute acoustic characteristics. For example, it is possible that the phoneme /eI/ would not be especially difficult to discriminate if its acoustic qualities were changed in some way so that it would be less similar to the other phonemes in the population or if it were judged in the context of a different population of sounds where it would be less confusable.

Detection task. No differences were observed in the detection data obtained from the two different experimental tapes. Hence data from all 16 subjects are combined. The results of the detection task are summarized in Table 2 in terms of mean latency for correct responses as a function of target level and target vowel. For the purposes of computing these means and for all subsequent analyses of this experiment, all latencies over 1 sec were truncated to 1 sec to eliminate very long reaction times resulting from failures of attention, and so on. In addition, all latencies over 4 sec were treated as errors to eliminate responses clearly made to the incorrect stimuli.[2] An analysis of variance computed on the mean latencies averaged across response item positions yielded 17 msec as the standard error of the entries of Table 2 for the present experiment. The targets involving the vowels /I/ and /eI/ showed longer latencies than the targets involving the other three vowels /i/, /æ/, and /aI/, $F(4, 56) = 6.23$, $p < .01$. Note that, as one might expect, the two vowels showing the longer latencies are just those found to be most difficult to identify according to the results of the identification task described above. The difference between latencies on phonemes and syllables was not significant, $F(1, 14) < 1$, nor was the interaction between target level and target vowel, $F(4, 56) = 1.64$, $p > .10$.

TABLE 2

MEAN LATENCY IN MSEC TO RESPOND IN DETECTION TASK OF EXPERIMENTS I AND II

Experiment	Target level	Target vowel /i/	/æ/	/aI/	/eI/	/I/
I	Phoneme	475	484	465	494	529
	Syllable	446	480	456	544	519
II	Phoneme	538	563	564	599	617
	Syllable	601	583	593	588	589

The error rate in the present experiment on lists with response items was relatively high, 10.6%. However, a speed-accuracy tradeoff cannot account for the results described above since the pattern of errors was consistent with the pattern of latencies on correct responses. More errors were made on the target vowels /eI/ (48) and /I/ (32), which had been shown to be relatively difficult to identify, than on the other three target vowels, /i/ (10), /æ/ (10), and /aI/ (2).

An additional analysis of variance was

[2] For the 960 subject-trials with response items, there were only four latencies over 4 sec, and 42 latencies between 1 and 4 sec, 24 with phoneme targets and 18 with syllable targets.

computed on the mean latencies averaged across target vowels rather than response item positions. Latencies decreased monotonically across response item position from 554 msec for Position 2, to 480 msec for Position 3, to 415 msec for Position 4, $F(2, 28) = 32.44$, $p < .01$. The monotonically decreasing function was found for both syllable and phoneme targets; the interaction of target level and response item position was not significant, $F(2, 28) < 1$. Unlike the vowel data, the effect of response item position may be accounted for in terms of a speed-accuracy tradeoff. For both phonemes and syllables, errors increased monotonically with increasing response item position, from 22 total errors for Position 2, to 37 for Position 3, to 43 for Position 4. This finding suggests that subjects' criteria to respond become more lax as the list progresses.

A further analysis of variance was computed on the mean latencies averaged across both target vowels and positions of the response item but separated in terms of the relationship of the level of the target and that of the response item (match vs. mismatch). Subjects responded faster when the levels of the target and the response item matched (460 msec) than when they mismatched (509 msec), $F(1, 14) = 9.71$, $p < .01$. The same pattern of results was found for phoneme and syllable targets. The interaction between the match-mismatch factor and target level was not significant, $F(1,14) < 1$. A speed-accuracy tradeoff is not involved in the present results since fewer errors were made on target items that matched the response items in level (42) than on target items that mismatched the response items (60). These results are consistent with those of McNeill and Lindig (1973), where responses to syllable targets were faster than those to phoneme targets when the response items were syllables; but responses to phoneme targets were faster than those to syllable targets when the response items were phonemes. The present results obtained with heterogeneous lists demonstrate that the matching factor found to be critical by McNeill and Lindig is not dependent on a homogeneous list construction. These results rule out the notion discussed above that the linguistic level at which speech perceivers focus their attention depends on the level of the search list as well as the level of the target. Within the conceptual scheme proposed above, these results suggest rather that the linguistic level at which speech perceivers focus their attention depends on the level of the *response item* as well as the level of the target. However, this suggestion seems dubious. How can the level of an item within a heterogeneous list determine the level at which that item is to be attended? In other words, if the subject knows the level of an item only *after* attention to that item, how can the level determine the nature of that attentional process?

A solution to this dilemma involves reconsideration of the implied assumption that the difference in linguistic levels is the only difference between a given syllable stimulus and the corresponding phoneme stimulus. If the vowel sounds in the syllable stimulus and the phoneme stimulus were physically different, one would not have to resort to the notion of linguistic levels to explain why matches were faster than mismatches. Subjects would naturally respond faster to a stimulus that was physically identical to the target than to one which was similar but not identical. Although the vowel sound in a given syllable stimulus and the vowel sound in the corresponding phoneme stimulus are physically identical for the first 200 msec, they differ for the last 125 msec. Hence only in the cases of a match are the stimuli physically identical for their complete duration. It is certain that the difference between the syllable and phoneme stimuli was even greater in the study by McNeill and Lindig (1973) where natural speech was employed, although a large difference for their stimuli at higher linguistic levels seems unlikely. If this explanation does hold, the linguistic level at which the speech perceiver focuses his attention cannot be

determined from either the present results or those of McNeill and Lindig.

Experiment II

Since Experiment I demonstrated that the matching factor is critical in determining the effect of target level on response latency even in heterogeneous lists, the present experiment was designed to eliminate this factor completely. This manipulation permits a comparison of the latencies to respond to phoneme and syllable targets that is free of bias. Specifically, heterogeneous lists were employed in the present experiment as in Experiment I; however, unlike Experiment I, in the present experiment the target and response item always matched in level. The response item to a phoneme target was always the phoneme itself, and the response item to a syllable target was always the syllable itself. No instances of either upward or downward search were included in the present experiment.

Method

Subjects and apparatus. Sixteen different Yale University undergraduates, who were native speakers of English with no known hearing defects, participated in this experiment for course credit. The subjects were divided into two groups of eight depending on their time of arrival for testing. The same apparatus was employed in the present experiment as in Experiment I.

Design and materials. The same stimuli were employed in the present experiment as earlier. The experimental tape constructed from these sounds for the identification task was the same as that employed in Experiment I. Since no difference between experimental tapes was found in the detection task of Experiment I, only one experimental tape was constructed for the present detection task; this tape was based on one of the two tapes employed in Experiment I. The practice tape constructed for the present experiment was similarly based on the practice tape employed in Experiment I. The targets for the two tapes used in the detection task were identical to those of the corresponding tapes of Experiment I but there were some changes in the list items. Whenever a response item did not match the target item on a trial in the tapes from Experment I, it was changed for the tapes of Experment II so that it would match the target item. The only other changes made were those necessary in order to maintain the other constraints employed in constructing the tapes for Experiment I.

Procedure. The procedure of the present experiment was analogous to that of Experiment I except for one change in the identification task and one change in the detection task, as described below.

The two groups of subjects differed only in the instructions they were given in the identification task. Subjects in Group 1 received instructions identical to those in Experiment I, whereas subjects in Group 2 received modified instructions. These subjects were given different names for the phoneme stimuli to insure that these stimuli were heard as phonemes and that the relationship of these stimuli to the other stimuli in the population was understood. In addition, these subjects were told to write on the response sheet for a given stimulus the number beside its name on the list of names rather than the name itself. This new procedure was used because the new names given the phoneme stimuli were quite lengthy: "the intial sound in *eat* and *eve*" (/i/), "the initial sound in *id* and *in*" (/I/), "the initial sound in *an* and *at*" (/æ/), "the initial sound in *aim* and *aid*" (/eI/), "the initial sound in *I'm* and *I've*" (/aI/).

The only change in the detection task from Experiment I to the present experiment was a simplification of the instructions. The subjects were told simply to press the response key during the presentation of the list as soon as they heard the target sound. No description of upward or downward search was included since no such trials were involved in the present experiment. Similarly, there was no need under the present conditions to tell the subject

to respond to the pronunciation not the spelling of the sound, so that this part of the instructions employed in Experiment I was also deleted from the present experiment.

Results and Discussion

Identification task. Although more errors were made on the identification task by subjects in Group 2 than subjects in Group 1, the same pattern of results held for the two groups. Hence in the discussion that follows the data from all 16 subjects are combined.

The results of the present identification task are summarized in Table 1 in terms of error percentages pooled across subjects as a function of vowel type and stimulus level. The present results essentially replicate those of Experiment I. Again the percentage of errors on the phonemes /I/ and /eI/ is significantly larger than on the other three phonemes, $T(10) = 6$, $p < .05$. Also the error percentages on the phonemes are less than on the corresponding syllables for the vowels /i/, /æ/, and /aI/, whereas the reverse holds for the vowels /eI/ and /I/.

Detection task. No differences were observed in the data of the detection task from Groups 1 and 2. Hence the data from both groups of subjects are combined in the following discussion.

The results of the detection task are summarized in Table 2 in terms of mean latency for correct responses as a function of target vowel. For computing these means, as in Experiment I, all latencies over 1 sec were truncated to 1 sec.[3] An analysis of variance computed on the mean latencies averaged across response item positions yielded 14 msec as the standard error of the entries of Table 2 for the present experiment. As in Experiment I, the vowels /I/ and /eI/, which had been shown in the identification task to be more difficult to identify, exhibited longer response latencies in the present task. The analysis revealed a significant effect of target vowel, $F(4, 56) = 3.97$, $p < .01$. In addition, although the factor of target level was not significant, $F(1, 14) = 2.79$, $p > .10$, the interaction of the factors target level and target vowel was significant, $F(4, 56) = 3.12$, $p < .05$. This interaction is quite interesting. Phoneme targets were responded to more quickly than syllable targets for the target vowels /i/, /æ/, and /aI/, but the reverse held for the vowels /I/ and /eI/. Identifiability may be able to account for the interaction completely since more errors were made on the identification tasks on phonemes than on syllables for the vowels /I/ and /eI/, but the reverse held for the other vowels, /i/, /æ/, and /aI/, just as latencies were longer in the detection task on phonemes than on syllables for those two vowels but the reverse held for the other three. On the basis of the present results one cannot state categorically that phonemes are detected more rapidly than syllables, as one would have expected if phonemes are perceptual units. However, neither can one state categorically that syllables are detected more rapidly than phonemes, as originally suggested by Savin and Bever (1970).

As in Experiment I, an additional analysis of variance was computed on the mean latencies from the present experiment averaged across target vowels rather than response item positions. Again latencies decreased monotonically across response item positions from 633 msec for Position 2, to 578 msec for Position 3, to 535 msec for Position 4, $F(2, 28) = 25.28$, $p < .01$. This trend is consistent with the notion that subjects' criteria to respond become more lax as the list progresses. Although a monotonically decreasing function was found for both phonemes and syllables, the factors of target level and response item position did interact significantly, $F(2, 28) = 4.15$, $p < .05$; the latency for phonemes was shorter than for syllables at Position 3 but was not different from that for syllables at Positions 2 and 4.

[3] For the 960 subject-trials with response items, no latencies were over 1.6 sec and only 24 latencies were between 1 and 1.6 sec, 15 with phoneme targets and 14 with syllable targets.

The error rate on lists including response items, 4.7%, was considerably lower in this experiment than in Experiment I. No regular pattern of errors was evident in the present data.

Summary and Conclusions

The aim of the present study was to consider the question posed by Savin and Bever (1970) whether syllables are detected more rapidly than phonemes. McNeill and Lindig (1973) suggested that the previous results showing that syllables are detected more rapidly than phonemes could be attributed to the following artifact of the previous experimental studies: The search lists were composed exclusively of syllables. McNeill and Lindig showed that the match or mismatch between the level of the target and the level of the search list was critical in determining the speed with which the target would be detected. The present study was designed to eliminate the critical matching factor of the previous studies by employing heterogeneous lists of phonemes and syllables. However, Experiment I demonstrated that the matching factor was not removed merely by changing the nature of the lists. The relative detection rates of phonemes and syllables were also shown to be affected by the match or mismatch between the level of the target and the level of the response item. Specifically, syllable targets were detected faster than phoneme targets when the response items were syllables, but phoneme targets were detected faster than syllable targets when the response items were phonemes. This result was explained by referring to the fact that although the vowel sounds in the syllable stimuli and the corresponding phoneme stimuli are physically identical for the initial 200 msec, they are not identical throughout their complete duration.

The second experiment of the present study was designed to remove the remaining biasing factor by including no instances of either upward or downward search; in every case the response item and the target item matched in level. Under these conditions it was discovered that the answer to the question whether syllables are detected more rapidly than phonemes is not a simple yes or no. Phonemes are detected more rapidly than syllables when the phonemes are relatively easy to identify, but syllables are detected more rapidly than phonemes when the phonemes are relatively difficult to identify.

Although the present results and those of the earlier studies do not allow one to conclude that linguistic level *per se* affects the speed with which speech sounds are detected, at least four factors have been pinpointed as critical in determining how fast a speech perceiver can detect a target sound: (*a*) the difficulty of comprehending the sentence in which the response item occurs (see for example, Foss, 1969; Foss & Lynch, 1969); (*b*) rhythmic cues in continuous speech (Shields, McHugh, & Martin, 1974); (*c*) the match or mismatch in linguistic level of target and response item (McNeill & Lindig, 1973; present study, Experiment I); and (*d*) the identifiability of the target sound (present study, Experiments I and II).

The present results are also pertinent to the more general question of the level at which speech perceivers enter the language hierarchy if two sets of assumptions are made: (*a*) The phonemic and syllabic levels are processed in discrete, nonoverlapping stages, and (*b*) any differences in response latencies reflect only differences in time for perceptual processing of the stimuli. The present findings are clearly incompatible with a model where phonemes are recognized before syllables. According to such a model vowel targets, which contain fewer phonemes than the corresponding vowel-consonant targets, would consistently show shorter latencies than vowel-consonant targets. However, in the present study vowel-consonant targets were detected before vowel targets in some instances The present results also seem incompatible with a model where syllables are recognized

before phonemes, but a firm conclusion cannot be made at present with respect to such a model. A model where syllables are recognized before phonemes would predict that syllable targets would be consistently detected before phoneme targets. This prediction seems incompatible with the results of the present study if the vowel targets are classified as phonemes and the vowel-consonant targets as syllables, in accordance with the classification scheme adopted for the purposes of the present study, since vowel targets were detected before vowel-consonant targets in some instances. However, it could be argued that vowel targets are classifiable as syllables as well as phonemes since they do stand alone. Under such a classification scheme, the present results would not be incompatible with a model where syllables are perceived before phonemes.

We have tentative support for two contradictory conclusions, that phonemes are not recognized before syllables and that syllables are not recognized before phonemes. One way to resolve this apparent contradiction is to revise the assumption of discrete, nonoverlapping processing stages by adopting a model involving concurrent rather than sequential processing of the phonemic and syllabic levels of speech, where the speech perceiver enters the linguistic hierarchy at the phonemic and syllabic levels more or less simultaneously. In other words, it seems best not to consider either the phoneme *or* the syllable as the basic perceptual unit but rather to consider the phoneme *and* the syllable as linguistic entities equally basic to speech perception.

Just as the assumption of discrete stages may need revision, the assumption that differences in latencies reflect only differences in perceptual processing time may need qualification (see also Foss & Swinney, 1973; Treisman & Squire, 1974). We hope to learn from detection tasks such as those employed in the present study when the various levels of linguistic analysis are carried out in speech perception. For example, we would like to conclude on the basis of the present results that in speech perception we do not necessarily process the phonemic level before the syllabic level. However, these conclusions about perception may not be justified since certain levels of analysis may be performed early in perception but may be difficult to access in a detection task. The phonemic level may be such a level as, most likely, is the level of the distinctive feature. Therefore, we are justified in generalizing our present results to speech perception only to the extent that the order of processing different linguistic levels for perception is the same as the order of accessing those levels in a detection task.

In conclusion, let us emphasize the methodological implications of the present study. These results make it clear that in any subsequent comparison of phonemes and syllables the identifiability of the stimuli must be considered. If identifiability were not controlled, no conclusion about relative detection rates of phonemes and syllables would be justified since a difference in either direction could be produced by suitable choice of stimuli. With one selection of stimuli, it would be possible to demonstrate that phonemes are responded to faster than syllables, whereas with another selection of stimuli, it would be possible to demonstrate that syllables are responded to faster than phonemes.

References

Bronstein, A. J. *The pronunciation of American English.* New York: Appleton-Century-Crofts, 1960.

Foss, D. J. Decision processes during sentence comprehension: effect of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 457–462

Foss, D. J., & Lynch, J. H. Jr. Decision processes during sentence comprehension: effects of surface structure on decision times. *Perception & Psychophysics*, 1969, **5**, 145–148.

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.

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

Miller, G. A. Decision units in the perception of speech. *IRE Transactions on Information Theory*, 1962, IT-8, 81–83.

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

Shields, J. L., McHugh, A., & Martin, J. G. Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, 1974, **102**, 250–255.

Treisman, A., & Squire, R. Listening to speech at two levels at once. *Quarterly Journal of Experimental Psychology*, 1974, **26**, 82–97.

(Received August 22, 1975)