Journal of Experimental Psychology
1974, Vol. 102, No. 2, 199–208

PERCEPTUAL UNITS IN SPEECH RECOGNITION [1]

DOMINIC W. MASSARO [2]

University of Wisconsin—Madison

The size of the sound stimulus employed in the first stage of speech processing was investigated in an attempt to determine the perceptual unit of analysis in speech recognition. It is assumed that the perceptual unit is held in a preperceptual auditory image until its sound pattern is complete and recognition has occurred. Vowels and consonant–vowel syllables were employed as test items in a recognition-masking task. The results show that recognition performance improved up to 200–250 msec. after presentation of the speech sound. The results were interpreted as evidence that the preperceptual auditory storage and perceptual processing of a speech sound does not exceed 250 msec., implying that some transformation of the speech signal must occur about every ¼ sec. Since the stimulus within this time period must function as a perceptual unit, perceptual units appear to be of roughly syllabic length.

The primary purpose of this study was to determine the size of the sound stimulus employed in the first stage of speech recognition. A listener has recognized (identified) a stimulus when he has determined that one of a possible set of alternatives was presented. Recognition of a stimulus is possible only if the information in the stimulus is sufficient to distinguish that stimulus from other possible stimulus alternatives. Recognizing speech continuously implies that each small portion of the sound-wave pattern uniquely determines a stimulus alternative. However, small portions of the acoustic input are not unique; there is no one-to-one mapping of stimulus to percept. Since the sound pattern must contain enough information for a consistent stimulus–percept mapping, larger chunks of the acoustic input are necessary for recognition.

Since the sound stimulus for speech recognition extends over time, the first part of the pattern must be held in some preperceptual form until the pattern is complete and recognition has occurred. Accordingly, it is assumed that the information in the sound pattern is held in a preperceptual auditory image and that the recognition process involves a readout of the information in that image. The duration of the preperceptual auditory image places an upper limit on the sound patterns that can be employed in the recognition process. Accordingly, to understand speech recognition, it is necessary to determine the maximum duration that information can be held in a preperceptual auditory image.

The minimal sound patterns that are usually recognized in continuous speech are referred to as perceptual units. In terms of the present analysis, the perceptual unit cannot exceed the duration of the image and must uniquely determine the appropriate perceptual response. The perceptual unit of analysis gives the perceptual system acoustic information that can be reliably correlated with information in long-term memory. The information in the unit is defined by a set of features that correspond to a list of features in long-term memory. The features made available by the stimulus are acoustic, whereas the list of features in memory is abstract. The recognition process must find a match between the acoustic features in the stimulus and a list of features in memory. The nature of perceptual units also provides information about the structure of long-term memory. A set of features in the acoustic signal indicates that the signal can be identified and can therefore be used to

[1] This investigation was supported in part by U.S. Public Health Service Grant MH-19399-02. A report of this research was presented at the Indiana Conference on Theoretical and Mathematical Psychology, Bloomington, April 1972. The reviewer provided a number of helpful comments on the manuscript.

[2] Requests for reprints should be sent to Dominic W. Massaro, Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706.

integrate with other units so that meaning can be derived from the message.

The following studies were carried out to determine the size of perceptual units employed in speech perception. It is necessary, therefore, to determine the temporal course of processing speech stimuli and to estimate the effective duration of the preperceptual image. Vowels and consonant–vowel syllables were employed as test items in an auditory recognition masking task (Massaro, 1970b). In this task, a test stimulus is presented, followed by a masking stimulus that occurs after a variable silent interstimulus interval. In a typical experiment, the test signals are short tones differing in pitch, and the listener's task is to identify the higher tone as high and the lower tone as low. The masking tone is presented at another frequency and at the same loudness as the test tones. The results indicate that recognition performance improves with increases in the silent intertone interval, up to about 250 msec. (Massaro, 1970b, 1971, 1972c).

These results provide information about the preperceptual auditory image of the test tone and the vulnerability of the auditory image to new inputs. Given that the test tone lasted only 20 msec., some preperceptual image must have remained for the perceptual processing necessary to improve recognition performance with increases in the silent intertone interval. This improvement in recognition also indicates that the masking tone terminated perceptual processing of the image. Since recognition performance leveled off at about 250 msec., the image effectively decayed within this period. Using this paradigm with speech stimuli should make it possible to determine the duration of preperceptual images and the processing time required for speech stimuli.

Experiment I

The first experiment demonstrates that the recognition of a short speech stimulus is not immediate, but rather requires time for perceptual processing. Perceptual processing refers to the analysis of physical features in the sensory input in order to recognize the stimulus. A measure of perceptual-processing time might be found in the durations of vowels in normal speech which are in the range of 150–250 msec. (Fletcher, 1953; House, 1961; Peterson & Lehiste, 1960). Since the pattern of sound-pressure fluctuations in a steady-state vowel repeats at the speaker's fundamental frequency, the extended duration of the vowel might be necessary for processing the information available in the vowel presentation. Vowels presented for very short durations can be identified if followed by a silent retroactive interval (Gray, 1942; Suen & Beddoes, 1972). However, if processing is interfered with by following the short test-vowel presentation with another speech sound, the test vowel should not be identified. This result would provide evidence that the duration of the vowel in normal speech allows time for processing, since the extended duration of the vowel protects it from later speech until recognition has been completed.

The *S*s were required to recognize a test vowel. On each trial, 1 of 2 vowels could be presented with equal probability. A second vowel, referred to as the masking vowel, followed the test vowel after a variable silent intervowel interval. It was assumed that the perceptual processing necessary for recognition of the test vowel could take place during the test-vowel presentation and the silent interval afterwards, but *not* during the masking-vowel presentation.

Method

Subjects. Three young adults from the University of Wisconsin community were employed as *S*s.

Procedure. The vowels of a male speaker were first recorded at the same fundamental frequency and amplitude. A steady-state segment of each vowel was stored digitally by a computer-controlled analog-to-digital converter. During this experiment, the vowel segment was played back, using a digital-to-analog converter. In the recognition masking task, the vowels /i/ as in *heat* and /I/ as in *hit* were employed as test items. The duration of the test vowel was 20 msec. The silent intervowel interval lasted 0, 20, 40, 80, 160, 250, 350, or 500 msec. The masking stimulus was a 270-msec. nonsense vowel made up of 2 alternating vowel segments. These 2 segments were taken from the

vowels /a/ as in *hat* and /U/as in *put* and lasted 45 msec. each.

The *S*s were tested simultaneously in a sound-insulated chamber. All experimental events were controlled by a PDP-8 computer. The vowels were presented binaurally over matched headphones (Grason-Stadler Model TDH-39) at a normal listening intensity (about 80-db. SPL). An *S* recorded his decision by pressing 1 of 2 push buttons, labeled i and I, respectively. Following the 2.5-sec. response period, feedback was given by illuminating a small light for 500 msec. above the correct response button. The intertrial interval was 2 sec.

On every trial, *S* heard a test vowel, followed in turn by a variable silent interval and the masking vowel. He identified the test vowel as i or I and was then informed of the correct answer for that trial. The *S*s had 1 day of practice identifying the test vowels without a masking vowel and 2 days of practice with a masking vowel before the experiment itself, which consisted of 2 days of 3 sessions each. There were 400 trials/session, and *S*s did not respond to the first 5 trials of each session. Since all experimental conditions were completely random, memory and decision factors were reasonably constant. The results, then, indicate the temporal course of the perceptual or identification process.

Results

The results of the experiment, presented in Figure 1, show that for each *S*, recognition improved from near chance to almost perfect performance with increases in the silent intervowel interval. The results indicate that an auditory image remained after the first vowel and was processed during the intervowel interval for recognition. The second vowel was effective in terminating the perceptual processing of the image. Figure 1 shows that the processing time necessary for asymptotic recognition differed for the 3 *S*s. Two *S*s (KB and MF) required about 180 msec. for optimal recognition. In contrast, the third *S* (DC) was able to identify the test vowel with only 80 msec. of additional processing time after its presentation.

These results do not necessarily estimate the duration of the preperceptual image or the perceptual processing time in typical vowel identification. Since performance reached asymptote at essentially perfect recognition, the recognition process could have been complete before the image decayed. To estimate the duration of the image, recognition accuracy must reach

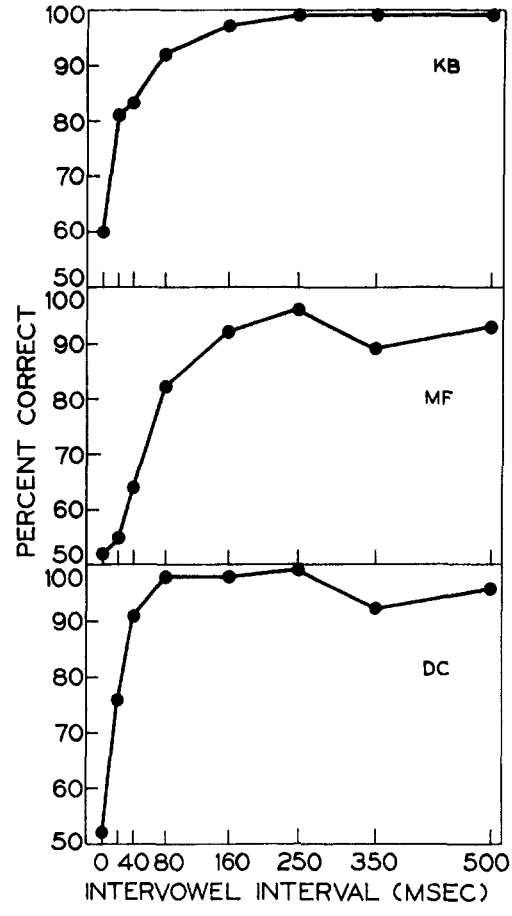

FIGURE 1. Percentage of correct test-vowel identifications for *S*s KB, MF, and DC as a function of the duration of the silent intervowel interval: Experiment I.

asymptote below perfect performance. In the present study, the test vowel was 1 of only 2 possible vowels. With only 2 alternatives, *S* can probably make his decision very rapidly, since fewer acoustic features of the vowels need be processed for accurate recognition. With a larger number of alternatives, however, *S* would have to process more features before reaching a decision. Accordingly, increasing the number of alternatives in this task should increase the effective perceptual-processing time up to the duration of the auditory image. Accordingly, Experiment II was carried out with 4 test alternatives in the recognition masking task.

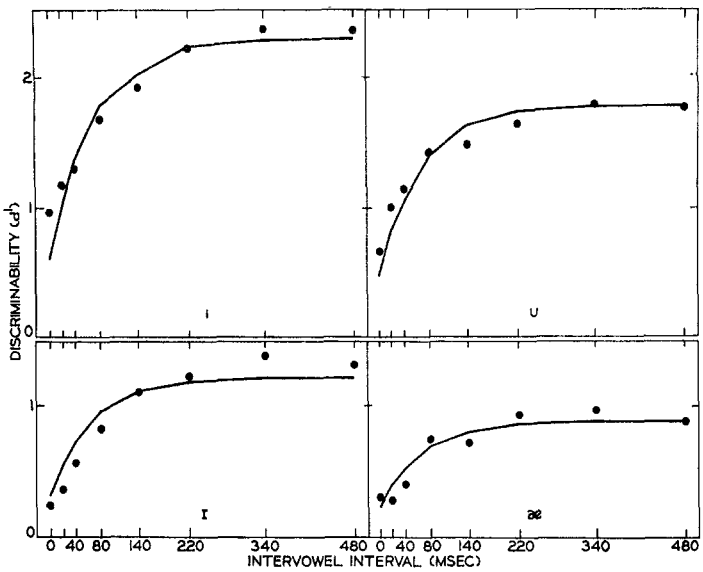

FIGURE 2. Discriminability values for the 4 test vowels as a function of the silent intervowel interval: Experiment II. (The lines are drawn through the predicted points given by Equation 1.)

EXPERIMENT II

Method

Subjects. Twelve University of Wisconsin undergraduates fulfilling a course requirement were tested for 5 days.

Procedure. In Experiment II, 4 vowels recorded by a female speaker were employed as test items. The fundamental frequency of the vowels was 192 Hz. The test alternatives were the vowels /i/ as in *heat*, /I/ as in *hit*, /a/ as in *hat*, and /U/ as in *put*. The duration of the test vowel was 20 msec. The silent intervowel interval lasted 0, 20, 40, 80, 140, 220, 340, or 480 msec. The masking stimulus was also chosen randomly from the 4 vowels, /i/, /I/, /a/, and /U/, and also lasted 20 msec.

Four *S*s were tested simultaneously in separate sound-attenuated rooms. All experimental events were controlled by a PDP-8L computer. The vowels were presented binaurally over matched headphones (Grason-Stadler TDH-49) at between 58- and 67-db. SPL. Each *S* recorded his decision by pressing 1 of 4 push buttons, labeled HEAT 1, HIT 2, HAT 3, and PUT 4, respectively. Following the 2.5-sec. response period, feedback was given by presenting the digits 1, 2, 3, or 4 on a visual display (IEE readout) for 500 msec. The intertrial interval was 2 sec. There were 2 300-trial sessions on each of 5 days. On the first day, *S* identified the test vowel without the masking vowel present. On Days 2, 3, 4, and 5 the experiment proper was carried out. Presentation of the 4 test alternatives, the 4 masking alternatives, and the 8 intervowel intervals was completely random, and each condition was programmed to occur equally often. Day 2 was treated as practice, so that the results are taken from the last 3 experimental days.

Results

The results were analyzed in terms of a signal detectability framework to provide a measure of vowel discriminability, d', which would be relatively independent of decision biases. It is possible to derive a measure of recognition performance for each of the 4 test vowels. The probability of identifying a vowel correctly is designated a hit, and responding with that vowel alternative to any other test alternative is designated a false alarm. For example, the probability of responding i given the test vowel /i/, $p(\text{i}|\text{i})$, would be a hit, whereas the probability of responding i given any other alternative, p $(\text{i}|\bar{\text{i}})$, would be a false alarm. The d' measure derived from these probabilities would be the index of discriminability for the vowel /i/. The d' values were obtained from the hit and false alarm rates (averaged across *S*s) in Elliot's (1964) tables.

Figure 2 presents the d' values for each test vowel as a function of the intervowel

interval. Identification of each vowel improved with increases in the silent interval between the test and masking vowels. These results can be used to estimate the duration of the preperceptual image and perceptual-processing time, since recognition reached asymptote below perfect performance. Since performance did not improve beyond an intervowel interval of 220 msec., the image could have decayed in this period. Figure 2 also shows that the test vowels were not recognized equally well, implying that the alternatives differed in discriminability. The rank ordering from least to most discriminable was /a/, /I/, /U/, and /i/.

The continuous lines in Figure 2 are drawn through the predicted points given by a perceptual-processing model (Massaro, 1970a), which assumes that readout of the information in the preperceptual image increases in a negatively accelerating manner with increases in processing time. More specifically, the perceptual strength of an item, as measured by its discriminability, d', follows an exponential growth function of time:

$$d' = \alpha(1 - e^{-\theta t}). \qquad [1]$$

The measure d' is the discriminability of the item after a presentation time of t sec. In the model, presentation time includes both the duration of the test item and the silent interval before the onset of the masking stimulus. Equation 1 indicates that the perception of a test item approaches an asymptote α at a rate of θ.

Each test vowel is assumed to function as a perceptual unit and therefore has a number of distinctive features that correspond to a feature list in long-term memory. The rate at which these features are processed is reflected in the value of θ. A vowel with a large number of distinctive features would be expected to have a large α value. A noisy or unclear vowel would have few distinctive features and a small α value. Thus, the α value can be thought of as an index of the number of discriminable features of a vowel.

According to the model, the overall differences found in identifying the test vowels should reflect differences in α, the measure of discriminability. However, θ, the rate of processing the information in the vowel, should be independent of α. In all cases, the number of features processed in any unit of time should be a constant proportion of the number of features that remain unprocessed. Equation 1 was applied to the observed results in Figure 2 by estimating a different parameter value of α for each test vowel and a single value of θ. As can be seen in Figure 2, the model describes the results fairly accurately. Thirty data points are predicted with only 5 parameter estimates. The parameter estimates of the α values for /i/, /I/, /a/, and /U/ were 2.29, 1.31, .88, and 1.77, respectively. The estimated value of θ was 15.17.

Table 1 presents the average recognition performance for the different test stimuli as a function of the masking vowel. As can be seen in the table, all masking vowels were equally effective in terminating perceptual processing of the test vowel. Of special interest is the finding that a vowel can mask itself. This means that S did not discriminate trials on which the test and mask were the same vowel. Since identification of the test vowel requires perceptual-processing time, it cannot be accurately compared to the masking vowel until it is recognized. The masking vowel overwrites the test vowel effectively, leaving very little information about the test-vowel presentation.

The results show that perception of a short vowel presentation improves with

TABLE 1

AVERAGE PERFORMANCE (MEASURED IN d' VALUES) FOR TEST VOWELS AS A FUNCTION OF MASKING VOWELS

Test stimulus	Masking stimulus			
	/i/	/I/	/a/	/U/
/i/	1.76	1.78	1.78	1.62
/I/	.78	.95	.96	.78
/a/	.56	.66	.23	.66
/U/	1.15	1.33	1.22	1.75
Average	1.06	1.18	1.05	1.20

increases in the silent interval after its presentation. This result provides evidence for the assumption that the duration of vowels in normal speech provides time for perceptual processing. However, increasing the duration of a vowel could also increase the amount of information in the vowel presentation. Increasing the duration of a vowel increases the information in the stimulus if one analyzes the frequency–amplitude spectrum of the steady-state vowel. With a short vowel stimulus, the distributions of energy around the respective formants are fairly large, and increasing the duration of the vowel would decrease the variance of the distributions. Accordingly, to measure the relative contribution of processing time and stimulus information in vowel perception, perception of a short vowel followed by a silent interval was compared to perception of that same vowel left on for the processing interval before presentation of the masking vowel.

Experiment III

Method

Subjects. Eight undergraduates fulfilling a course requirement at the University of Wisconsin served as *S*s for 5 days.

Procedure. The 4 test alternatives used in Experiment II were also employed in this study. The vowels were first modified digitally to give exactly the same steady-state loudness (65-db. SPL). Since the vowels had a fundamental frequency of 192 Hz., 5 fundamental segments gave a stimulus duration of 26 msec. In the continuous-processing conditions, vowels had to be presented at durations that exceeded our computer storage capacity. Accordingly, we stored 26 msec. of each vowel and repeated this segment until the desired duration was presented. The masking vowel was a 208-msec. continuous nonsense vowel made up of 2 repetitions of the 4 test vowels, /i/, /I/, /a/, and /U/, played for 26 msec. each.

On the first day of the experiment, *S*s listened to the test vowels played at durations of 208, 104, and 52 msec., respectively, in 3 successive 300-trial sessions. Feedback was given simultaneously with the test-vowel presentation. The experiment proper was carried out on the following 4 days. In the silent-processing condition, the test vowel was presented for 26 msec., followed in turn by a silent interval (0, 26, 52, 78, 130, 182, 260, or 390 msec.) and the masking vowel. In the continuous-processing condition, the test vowel was presented for 26, 52, 78, 104, 156, 208, 286, or 416 msec., followed immediately by the masking vowel. All experimental conditions (4 Test Vowels × 2 Processing Conditions × 8 Processing Intervals) were completely random in a given session and were programmed to occur equally often. Therefore, on every trial, *S* heard a test vowel of variable duration, followed in turn by a variable silent interval and a masking vowel. Following a 2-sec. response period, feedback was presented. The intertrial interval was 1 sec. All other procedural details were similar to Experiment II. Two 300-trial sessions were given, and the first 5 trials on each day were ignored in the data analysis.

Results

The results shown in Figures 3, 4, 5, and 6 present discriminability values for each vowel as a function of processing time and the silent and continuous conditions. For every vowel, identification was better when the vowel was left on for the processing interval than when a 26-msec. vowel was followed by a silent-processing interval. More specifically, for all test vowels, performance reaches asymptote at a higher level of performance for the continuous- than for the silent-processing condition, suggesting that there was more stimulus information in the test vowel in the former than in the latter processing condition. Furthermore, the results also indicate that performance reaches asymptote at the same processing interval for the silent- and continuous-processing conditions. Accordingly, it appears that the rate of processing the information in the test vowel did not differ under the 2 processing conditions.

This interpretation of the results can be described in the framework of the perceptual-processing model. We first assume that a 52-msec. vowel produces more information than a 26-msec. vowel, but that the information in a vowel presentation does not increase with increases in duration beyond 52 msec. It follows that the continuous condition should only have one value of α at test vowel durations of 52 msec. or more. We assume that this α value is some constant K, times the α value for the corresponding 26-msec. vowel presentation. Accordingly, to obtain the theoretical predictions, it is necessary to estimate a different α for each of the 4 test vowels, the constant K, and a value

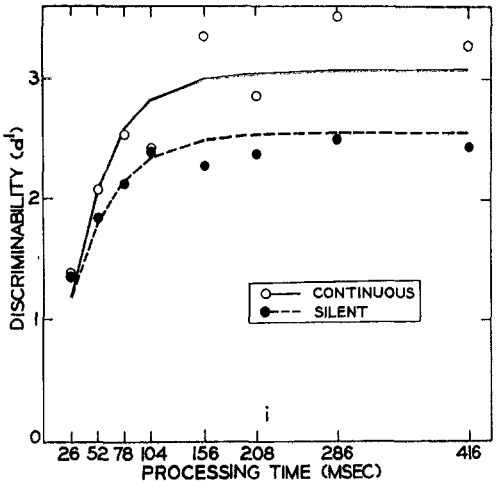

FIGURE 3. Discriminability values for the test vowel /i/ as a function of processing time during the continuous vowel presentation or the silent interval after a 26-msec. vowel presentation: Experiment III. (The lines are drawn through the predicted points given by Equation 1.)

for θ, the rate of perceptual processing, which should be constant under both the silent and continuous tone conditions.

The predicted curves are given by the continuous lines in Figures 3, 4, 5, and 6. The model does a reasonably good job of describing the data, considering the fact that 64 points are predicted by estimating only 6 parameters. The parameter values for α were 2.57, 1.41, 1.24, and 1.91 for the vowels /i/, /I/, /a/, and /U/, respectively. The estimated value for K was 1.20, and the value for θ was 23.77.

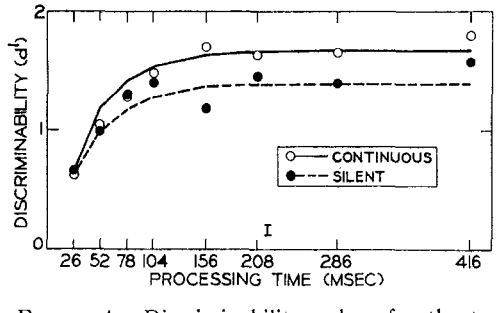

FIGURE 4. Discriminability values for the test vowel /I/ as a function of processing time during the continuous vowel presentation or the silent interval after a 26-msec. vowel presentation: Experiment III. (The lines are drawn through the predicted points given by Equation 1.)

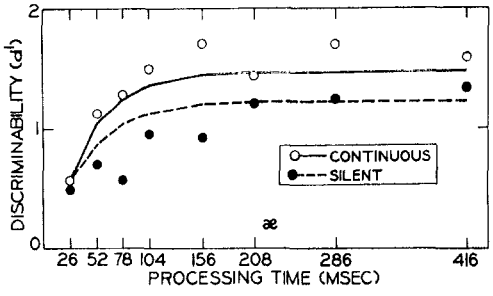

FIGURE 5. Discriminability values for the test vowel /a/ as a function of processing time during the continuous vowel presentation or the silent interval after a 26-msec. vowel presentation: Experiment III. (The lines are drawn through the predicted points given by Equation 1.)

In the silent-processing condition, recognition performance improved with increases in the silent interval, leveling off after roughly 200 msec. It might therefore be concluded that the preperceptual image of the test vowel decayed within this period. However, recognition performance reached asymptote at the same time in the continuous vowel condition. In this case, the image should not have decayed, since the test vowel was still being presented. It appears that recognition leveled off because S processed all the information available in the stimulus in the first 200 msec. In terms of the perceptual-processing model presented here, the rate of processing was

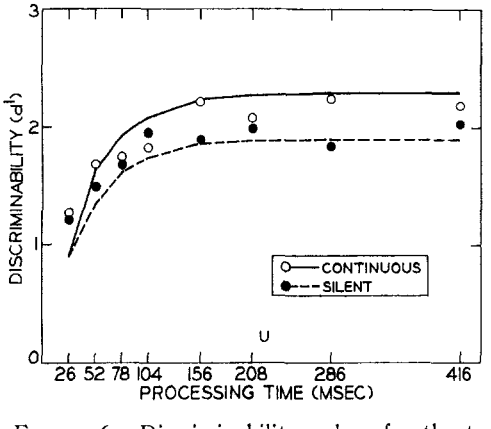

FIGURE 6. Discriminability values for the test vowel /U/ as a function of processing time during the continuous vowel presentation or the silent interval after a 26-msec. vowel presentation: Experiment III. (The lines are drawn through the predicted points given by Equation 1.)

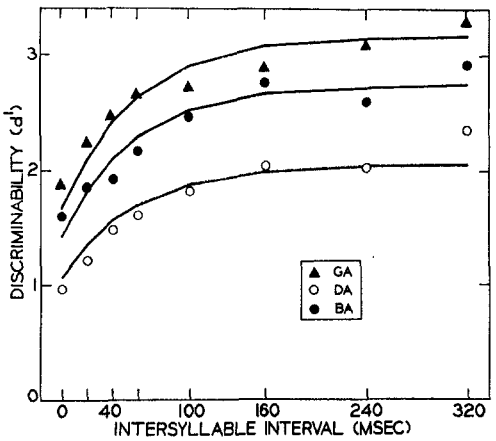

FIGURE 7. Discriminability values for the 3 test syllables as a function of the silent intersyllable interval: Experiment IV. (The lines are drawn through the predicted points given by Equation 1.)

large enough that performance reached asympote within 200 msec. It is possible that further increases in the number of test alternatives will differentially enhance performance in the continuous- relative to the silent-processing condition.

The results indicate that perceptual processing of a continuous vowel presentation continues for about 200 msec. This processing is terminated by a masking-vowel presentation. These results support the idea that the extended duration of the vowel in continuous speech allows time for perceptual processing, so that the vowel can be identified. The processing time during the steady-state vowels could also provide time for recognition of some consonants. The stop consonants are characterized by rapid transitions in the formants of the acoustic signal (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). The stop consonants /g/, /d/, and /b/ can be distinguished from each other by the second formant transition, which varies as a function of place of articulation (Delattre, Liberman, & Cooper, 1955). It has been argued that the stop consonant could not be recognized while it is presented, but would require some silent time or steady-state period for recognition to take place (Massaro, 1972b).

This analysis implies that the stop consonant–vowel (CV) transition would function as a perceptual unit in speech perception. If the vowel is integrated with the CV transition, the extended duration of the vowel in normal speech would provide time for perceptual processing. Accordingly, the first part of the CV pattern probably could be identified if it is presented alone and followed by a silent interval. However, if the short CV syllable were followed by a speech sound that could not be integrated with it, perception should be disrupted, and backward recognition masking should occur. To test this, we took the CV transitions of 3 stop CV syllables and used these as test items in our backward-masking experiment.

EXPERIMENT IV

Method

Subjects. Six University of Wisconsin undergraduates fulfilling a course requirement were tested for 5 days.

Procedure. Three CVs were employed as both test and masking stimuli. The CVs were BA, DA, and GA (A as in *box*). The 2 formant CVs were synthesized at Haskins Laboratory[8] and recorded digitally for playback during the experiment. Enough of the CV syllable was presented for it to sound speechlike. Each CV syllable was of 42-msec. duration (30 msec. of transition plus 12 msec. of steady-state vowel) and was presented at 76-db. SPL. On each trial, 1 of the 3 CV syllables was presented, followed in turn by a variable interval and 1 of the 3 syllables. The silent intersyllable interval lasted 0, 20, 40, 60, 100, 160, 240, or 320 msec. Each *S* recorded his decision by pressing 1 of 3 push buttons labeled GA 1, DA 2, and BA 3, respectively. Following the 1.5-sec. response interval, feedback was provided by visually presenting the digits 1, 2, or 3 for 500 msec. The intertrial interval was 1.5 sec. Presentation of the 3 test alternatives, 3 masking alternatives, and 8 intersyllable intervals was completely random. The alternative GA could occur 2 out of 8 times and BA and DA each occurred 3 out of 8 times. All other procedural details were the same as in Experiment II.

On the first day, *S*s identified the test syllable without the masking syllable present for 300 trials. The masking stimulus was presented in the second session. On Days 2, 3, 4, and 5, the experiment proper was carried out. There were 2 300-trial sessions per day. Only the results of Days 3, 4, and 5 were included in the data analysis.

[8] David B. Pisoni kindly provided a recording of the original unshortened syllables.

Results

The results, presented as in Figure 7, show that identification performance improved with increases in the silent intersyllable interval for all 3 CVs. The continuous lines are drawn through the predicted points given by Equation 1. Four parameter values were estimated: an α value for each CV test alternative, and a value of θ. The parameter estimates were 3.18, 2.06, and 2.75 for the values for GA, DA, and BA, respectively. The parameter estimate for the rate of perceptual processing, θ, was 17.28. Table 2 shows that the similarity between the test and masking stimuli did not systematically affect identification performance. However, BA produced less overall interference than GA or DA as a masking stimulus. This result might reflect the fact that the low frequencies of the syllable BA are less effective in terminating perceptual processing of an earlier CV presentation.

TABLE 2

AVERAGE RECOGNITION PERFORMANCE (MEASURED IN d' VALUES) FOR TEST SYLLABLES AS A FUNCTION OF MASKING SYLLABLES

Test syllable	Masking syllable		
	GA	DA	BA
GA	3.13	2.48	2.32
DA	1.40	1.58	1.94
BA	1.66	2.38	3.18
Average	2.10	2.22	2.72

DISCUSSION

These studies were carried out to determine the size of perceptual units employed in speech perception. It was argued that the perceptual unit could not be larger than the temporal life of a preperceptual image of the speech stimulus. If the information in the image decays before it is synthesized, recognition would not be possible. The results indicate that processing the speech stimuli employed here occurs within 200–250 msec. after the stimulus presentation. Accordingly, some transformation of the speech signal must occur about every $\frac{1}{4}$ sec. The stimulus within this time period must function as a perceptual unit. There must be sufficient information in this signal (and in the context in normal speech) for contact with information in long-term memory. To the extent the present studies simulate the processing of real speech, it might be concluded that the first stage of speech recognition must operate on acoustic segments that are of less than 250-msec. duration. If this is the case, perceptual units appear to be of roughly syllabic length, as suggested by Huggins (1964) and Massaro (1972b).

There are at least 2 apparent difficulties with assuming that the first stage of recognition involves a transformation or synthesis within 250 msec. after a sound pattern is presented. The first is that the utilization of auditory information occurs across much longer temporal periods. The memory for the quality of a speaker's voice requires an auditory memory that can last indefinitely. The interpretation of prosodic information such as stress requires an auditory memory that lasts for the length of a phrase or sentence. Most importantly, the recognition of sound patterns depends on an auditory memory for those sound patterns or even other sound patterns by the same speaker. For example, Ladefoged and Broadbent (1957) showed that the acoustic characteristics of some vowels of a speaker influenced the recognition of other vowels by the same speaker. Accordingly, the listener must have remembered some of the auditory characteristics of the vowels for longer than 250 msec. The auditory memory illustrated by these examples, however, is not preperceptual, but is auditory information that has been actively synthesized.

Massaro (1972a, 1972b) has distinguished between two stages of auditory information processing which involve preperceptual and synthesized auditory storage, respectively. The first stage involves recognition of a sound pattern held in a preperceptual auditory image. The recognition of the sound pattern produces a synthesized auditory percept in synthesized auditory storage. Synthesized auditory storage has a much longer life span than preperceptual images and could be responsible for auditory memory effects that operate over longer time periods than are given by preperceptual auditory storage. Accordingly, the auditory storage responsible for memory for voice quality, the Ladefoged and Broadbent (1957) effect, and for interpreting prosodic features of a sentence, is not preperceptual, but is at the level of synthesized auditory storage. (For a more complete discussion of the differences in pre-

perceptual and synthesized auditory storage, see Massaro, 1972a, 1972b.)

The second problem is that it is well known that the acoustic sound pattern corresponding to a syllable changes with changes in the speaker, the rate of speaking, the stress patterns, and (most importantly) the influence of neighboring syllables (Fant, 1962; Ohman, 1966). However, this problem is not unique to speech recognition, but is also the major deterrent to the development of a recognition device for handwritten text. This observation tells us that the representation of a perceptual unit in long-term memory must be a highly flexible one, with a number of normalization algorithms that operate on the sound-pattern input. Luckily, in normal speech, context or redundancy significantly reduces the number of valid alternatives for each sound pattern. The sound pattern can, therefore, be noisy, since redundancy reduces the number of acoustic features that are necessary to recognize the pattern correctly.

REFERENCES

DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. Acoustic loci and transition cues for consonants. *Journal of the Acoustical Society of America*, 1955, **27**, 769–773.

ELLIOT, P. B. Tables of d′. In J. A. Swets (Ed.), *Signal detection and recognition by human observers.* New York: Wiley, 1964.

FANT, C. G. M. Descriptive analysis of the acoustic aspects of speech. *Logos*, 1962, **5**, 3–17.

FLETCHER, H. *Speech and hearing in communications.* P.inceton, N.J.: D. Van Nostrand, 1953.

GRAY, G. W. Phonemic microtomy: The minimum duration of perceptible speech sounds. *Speech Monographs*, 1942, **9**, 75–90.

HOUSE, A. A. On vowel duration in English. *Journal of the Acoustical Society of America*, 1961, **33**, 1174–1178.

HUGGINS, A. W. F. Distortion of the temporal pattern of speech: Interruption and alternation. *Journal of the Acoustical Society of America*, 1964, **36**, 1055–1064.

LADEFOGED, P., & BROADBENT, D. E. Information conveyed by vowels. *Journal of the Acoustical Society of America*, 1957, **29**, 98–104.

LIBERMAN, A. M., COOPER, F. A., SHANKWEILER, D. P., & STUDDERT-KENNEDY, M. Perception of the speech code. *Psychological Review*, 1967, **74**, 431–461.

MASSARO, D. W. Perceptual processes and forgetting in memory tasks. *Psychological Review*, 1970, **77**, 557–567. (a)

MASSARO, D. W. Preperceptual auditory images. *Journal of Experimental Psychology*, 1970, **85**, 411–417. (b)

MASSARO, D. W. Effect of masking tone duration on preperceptual auditory images. *Journal of Experimental Psychology*, 1971, **87**, 146–148.

MASSARO, D. W. *Preperceptual and synthesized auditory storage.* (Wisconsin Studies in Human Information Processing Rep. 72-1) Madison: University of Wisconsin, Wisconsin Mathematical Psychology Program, November 1972. (a)

MASSARO, D. W. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 1972, **79**, 124–145. (b)

MASSARO, D. W. Stimulus information versus processing time in auditory pattern recognition. *Perception & Psychophysics*, 1972, **12**, 50–56. (c)

OHMAN, S. E. G. Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, 1966, **39**, 151–168.

PETERSON, G. E., & LEHISTE, I. Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 1960, **32**, 693–703.

SUEN, C. Y., & BEDDOES, M. P. Discrimination of vowel sounds of very short duration. *Perception & Psychophysics*, 1972, **11**, 417–419.

(Received January 13, 1973)