Perception & Psychophysics
1975, Vol. 18 (2), 149-157

Temporally segmented speech

A. W. F. HUGGINS
Research Laboratory of Electronics, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

Temporally segmented speech is continuous speech broken up by the insertion of silent intervals. The durations of the resulting speech intervals and silent intervals can be varied independently. When silent intervals are held constant at 200 msec, and speech interval duration is varied, intelligibility falls from about 90% to about 10% as speech interval duration is reduced from 200 to 30 msec. When speech interval duration is held constant at 63 msec, and silent interval duration is varied, intelligibility recovers from its asymptotic value of about 50% with long silent intervals, to 100% as the silent intervals are shortened from about 120 msec to about 60 msec. Implications for short-term acoustic storage are discussed.

Transformations of the speech wave that interfere drastically with its intelligibility have provided the starting points for many of the early studies of speech perception. The details of how the transformations have their effects are well understood for most of the common forms of distortion, but it is always exciting when a new transformation is discovered that unexpectedly disrupts the perceptual process, perhaps because such discoveries are becoming increasingly rare. Presumably, perceptual mechanisms for handling all naturally occurring transformations of the input signal have developed in response to evolutionary pressures. Therefore, transformations that do *not* occur in the natural world often provide the most significant insights, since they may be able to expose and explore "chinks" in the perceptual armor.

A prime example of such a transformation was reported by Cherry and Taylor (1954). A continuous speech message was switched back and forth between the listener's left and right ears, so that all the message entered his head, but never through left and right ears simultaneously. Intelligibility was drastically reduced at a switching rate of about 3 Hz, although it was relatively unaffected by rates a few times faster or slower. Cherry and Taylor's experiment was designed to measure the time required to switch attention from one ear to the other. Attention-switching time could be determined from the signal-switching rate that yielded the lowest intelligibility, they argued, since attention and signal were then exactly out of phase, with the attention reaching a given ear just as the signal left it. But this interpretation fails to account for their further result, that simply interrupting the signal produced a similar minimum of intelligibility, at the same cyclic rate. Under these conditions, no switching of attention between the ears is necessary.

A later study (Huggins, 1964) replicated Cherry's result, though the effect was less dramatic, and also showed that the critical rate of alternation, where intelligibility was lowest, varied with the playback speed of the speech. The latter result has recently been repeated by Wingfield and Wheale (Note 1). This finding argues against an interpretation such as Cherry offered, couched in terms of a temporal parameter of the perceptual apparatus, and implicates instead a temporal property of the *speech*. In a further result, Huggins showed that the intelligibility of the alternated speech could be quite accurately predicted, given some reasonable assumptions, from the intelligibilities of the two complementary interrupted messages in the left and right ears. This implies that preliminary processing of the interrupted speech is carried out separately for each ear, during alternated speech, perhaps because the two inputs are disparate enough to prevent perceptual fusion. Fusion is possible at a higher level *only* because both left- and right-ear interrupted signals were derived from a single message.

Both of the foregoing findings suggest that speech alternated at the critical rate is made less intelligible because the speech reaches each of the listener's ears in "packets," with each packet separated from its neighbors by silence. The alternation of the signal plays no part in the phenomenon, except that it arranges for the continuous speech to be presented to the two ears in packets. The foregoing argument suggests that a parallel interference with intelligibility might occur if a continuous speech signal is broken up into packets simply by the insertion of silent intervals. This was verified in the first experiment on temporally segmented speech (Huggins, 1972). The great advantage of temporally segmented speech, from the experimental point of view, is that the durations of the speech intervals (or packets) and those of the intervening silent intervals can be varied inde-

A preliminary report of this research was presented at the 83rd meeting of the Acoustical Society of America, Buffalo, 1972. The research was supported by NIH Grant NS04332. The author is now also at Bolt Beranek and Newman Inc.

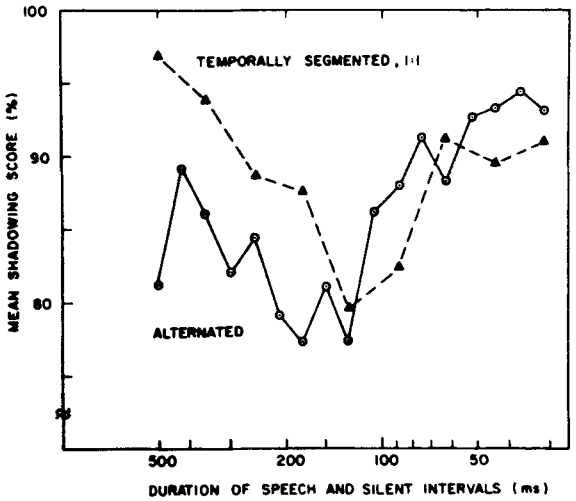

Figure 1. Shadowing performance is compared for alternated and temporally segmented speech. In alternated speech, successive speech intervals are presented alternately to the left and right ears, with the other ear receiving silence of the same duration (data from Huggins, 1964). In the temporally segmented speech, each speech interval is followed by a silent interval of the same duration (data from Huggins, 1972).

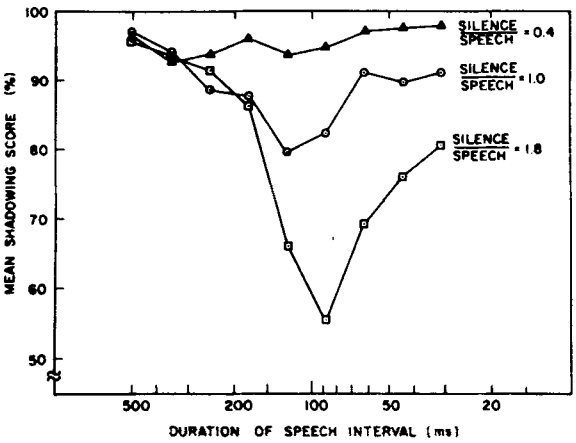

Figure 2. Shadowing performance is plotted as a function of *speech* interval duration, for three different versions of temporally segmented speech. The *same* speech intervals were separated, in the three versions, by silent intervals lasting 0.4, 1.0, and 1.8 times the duration of the adjacent speech intervals (data from Huggins, 1972).

pendently. Taking advantage of this fact, the same speech passages were temporally segmented three different ways, using a computer. Speech-interval durations were varied between 31 and 500 msec, and the speech intervals were identical across the three versions, with respect to both duration and content. The three versions differed only in the duration of the inserted silent intervals, which were related to those of the adjacent speech intervals by constant multipliers of 0.41, 1.0, and 1.83, respectively. The three versions will be referred to as those with "short silence," "equal silence," and "long silence." The data from this experiment are replotted in Figures 1, 2, and 3.

First, in Figure 1, the intelligibility function for the equal-silence version, in which speech and silent intervals were of the same duration, is compared with the corresponding function for alternated speech (Huggins, 1964), which also consists of speech and silent intervals of equal duration. ("Intelligibility" here refers to the percentage of words subjects were able to shadow.) The similarity of the two functions strongly supports the interpretation of the Cherry effect as a result not of the *alternation* of the signal between the ears, but rather of the signal reaching each ear in packets, each packet separated by silence from adjacent packets in the same ear. (The leftmost data point in the alternated function was strongly influenced by a learning effect that was inadequately counterbalanced.)

The intelligibility functions for the versions with short, equal, and long silent intervals are compared in Figure 2, where intelligibility is plotted against speech interval duration. Each function shows a V-shaped minimum, that becomes progressively deeper and occurs at progressively shorter speech intervals, across the short, equal, and long versions. The left-hand sides of the three functions are similar enough to suggest the possibility of a single underlying function. Thus, the decline of intelligibility as speech intervals are shortened seems to depend only on the duration of the speech intervals, and not the silent intervals. But the *right*-hand sides of the three functions in Figure 2 are *not* in agreement, and the only difference between the three stimulus tapes was the duration of the silent intervals. Perhaps the recovery in intelligibility on the right of the figure depends only on the silent intervals? To test this possibility, the intelligibility scores of the short-silence, equal-silence, and long-silence versions are plotted in Figure 3 as a function of silent-interval duration. This brings the right-hand sides of the three functions into rough agreement, and suggests that the recovery of intelligibility, when speech and silent

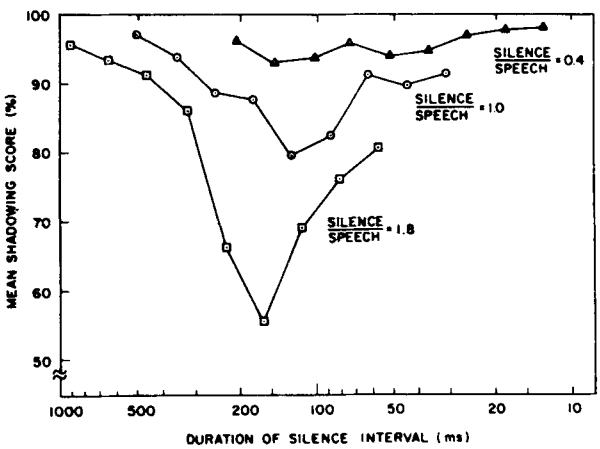

Figure 3. The data shown in Figure 2 are replotted, showing shadowing performance as a function of *silent*-interval duration for the three versions of temporally segmented speech.

intervals are further shortened, may depend mainly on the duration of the silent intervals, and less on the duration of the speech intervals.

One way of interpreting the foregoing findings, which underlies the design of the experiments described below, is as follows. When the silent intervals are long (data points to the left of the minima in Figure 2), each speech interval has to be processed as an isolated excerpt of speech, and the intelligibility of an excerpt decreases as it gets shorter. Secondly, as the silent intervals are progressively shortened, a point is reached where the ear begins to "bridge the gap" and relate acoustic events that occur after a silent interval to events that occurred before it. At this point, intelligibility begins to improve again.

In all the experiments described above, correlated changes were made in speech and silent intervals simultaneously. The foregoing interpretation of the results can best be tested by varying either speech or silent-interval duration, while holding the other constant. When the speech intervals are to be varied, the constant silent interval must be long enough that no "gap-bridging" can occur. This duration can be estimated by extrapolating down the common function supposedly underlying the right hand side of Figure 3, giving a value of about 200 msec. Similarly, when silent intervals are to be varied, speech intervals must be short enough that intelligibility is low in the absence of gap-bridging. Extrapolating down the left-hand sides of the functions in Figure 2 suggests that intelligibility will reach zero when the speech intervals are shortened to about 60 msec. Thus, if silent intervals are held constant at 200 msec, the intelligibility of the speech intervals can be measured directly, as a function of their duration. Similarly, if speech intervals are held constant at 60 msec, it should be possible to measure directly the ear's ability to bridge a silent interval, as a function of its duration.

METHOD

Two sets of nine 150-word passages were each temporally segmented in two ways, using a PDP-9 computer. In the "speech-varying" version, silent intervals were held constant at 200 msec and speech-interval duration increased in logarithmic steps from 31 msec in the first passage of each set to 500 msec in the ninth. In the silence-varying version, speech intervals were held constant at 63 msec and silent-interval duration increased in the same logarithmic steps from 31 msec in the first passage of each set to 500 msec in the ninth. Subjects heard the temporally segmented speech at a comfortable level (70 dB SPL) through noise-excluding headphones (Sharpe HA-660), and repeated as much as possible of the message into a microphone. The scoring method was explained in detail to the subject, and he was asked to try to maximize his score. He was told to shadow at whatever delay he found easiest, but a few words behind the input. He was warned against trying to speak in unison with the input, since most subjects find it too difficult and their "syllabic mutterings" (Cherry & Taylor, 1954) are impossible to score. They were also warned against storing up long strings to produce in a rush, because of the risk of losing the stored string upon hearing something unintelligible. Several minutes' practice was given in shadowing undegraded passages similar to the experimental passages, read by the same talker.

Eight subjects each shadowed the speech-varying version of one set of nine passages and the silence-varying version of the other set. Subjects were run individually, and the order of presentation and the two sets of passages were appropriately counterbalanced. The subject's shadowing responses were recorded on a second recorder, and a shadowing score was derived, for each condition, by counting the number of words correctly repeated from the middle 100 words of each 150 word passage. The first 35 and last 15 words of each passage were not scored, to exclude start-up difficulties and any recency effects.

RESULTS

The pooled shadowing scores are shown in Figure 4. When silent intervals are held constant at 200 msec and speech intervals are varied (Curve A), intelligibility declines from close to 100%, when speech intervals last 200 msec or more, to about 10%, when speech intervals last 31 msec. This is in excellent agreement with the hypothesis under test. Secondly, when speech-interval duration is held constant at about 63 msec (Curve B), silent intervals of 63 msec or less do not affect intelligibility significantly. But, as silent-interval duration is increased from 63 to 125 msec, intelligibility declines rapidly to about 55%, where it remains despite further increments in silence duration. The asymptotic part of Curve B was quite unexpected—but note that the asymptotic intelligibility of 55%, with long silent intervals, is exactly what would be predicted from Curve A for speech intervals of 63 msec. This provides striking confirmation of the hypothesis under test. Note further that the 200-msec silent intervals chosen to temporally segment the

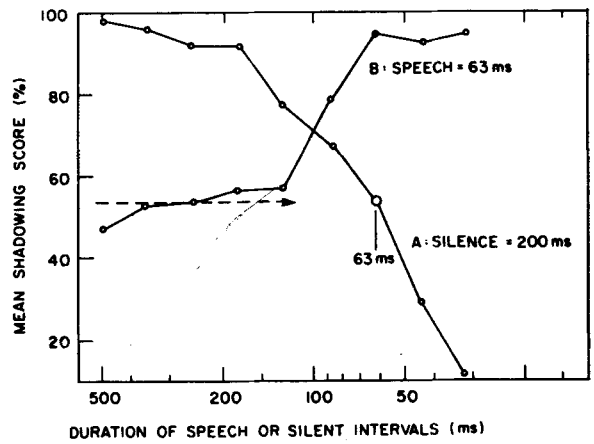

Figure 4. Shadowing performance for temporally segmented speech with speech and silent intervals varied independently. The function marked "A" represents the data when *speech* intervals were varied and silent intervals were held constant at 200 msec. The function marked "B" represents the data when *silent* intervals were varied and speech intervals were held constant at 63 msec (gap-bridging). Note: The asymptotic part of Curve B, for long silent intervals, is at just the value one would predict from Curve A for speech intervals of 63 msec.

speech-varying version are long enough to lie on the asymptotic part of Curve B, thus confirming that the speech intervals were effectively presented in isolation.

Clearly, then, the intelligibility of temporally segmented speech depends upon the durations of both the speech and silent intervals. To test the adequacy of the proposed explanation, an attempt was made to use the results of the present experiment to predict the intelligibility functions shown in Figure 2. The model proposes that the probability of correct recognition of the speech in a speech interval (P_S) is equal to the probability that it can be recognized in isolation (P_A), plus the probability that it can be combined with adjacent speech intervals across the intervening silent intervals (P_B), if it was *not* recognised in isolation. That is to say,

$$P_S = P_A + (1 - P_A) \cdot P_B.$$

Values for P_A as a function of speech-interval duration were read directly from Curve A in Figure 4. Corresponding values for P_B as a function of silent-interval duration were derived from Curve B by correcting for the asymptotic intelligibility of 55%. That is, corrected score = (observed score - 55)/(100 - 55). In Figure 5, the predicted values of P_S are compared with the observed values for the equal-silence version (silence/speech = 1.0) and for the long-silence version (silence/speech = 1.82). The agreement is remarkably good, given the simplicity of the model, but it should be remembered that the two

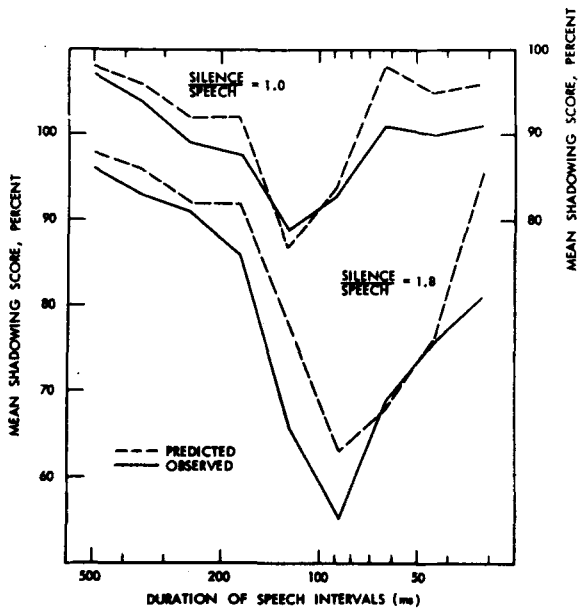

Figure 5. Observed shadowing performance for two of the three versions of temporally segmented speech, from Figures 2 and 3, is compared with results predicted by a simple model for combining the separate effects of speech- and silent-interval duration, derived from the data plotted in Figure 4. Full details are given in the text.

experiments had much more in common than usually occurs. The speech materials for both studies came from the same master tape, and therefore used the same talker. They were processed and presented through the same equipment, by the same experimenter, using the same procedure, to subjects drawn from the same student population. Only the temporal parameters of the segmented speech differed.

DISCUSSION

Before presenting possible interpretations of the results, and relating them to studies of short-term auditory storage, some objections to the present experiment need to be answered.

A Defense of Shadowing

There continues to be considerable criticism of shadowing as a task suitable for measuring intelligibility (e.g., most recently, Speaks & Trooien, 1974). Cherry and Taylor (1954) were aware of the problem, and were careful to stress that (1) their tests were essentially behavioristic, (2) the human subject was used only as a transducer, and (3) their results were stated in terms of "success" scores, not intelligibility scores. The main criticism is that the shadowing task interferes with the recognition task, and that subjects could recognize the speech perfectly well if they did not have to shadow it at the same time. The most obvious defense against this criticism is that even if it is true, it is beside the point. Subjects are not able to perform both recognition and shadowing tasks without error at the critical rates of alternation. However, at rates a few times faster or slower, they *can* perform both tasks, so the shadowing task apparently does not *always* interfere with the recognition task. Thus, the degraded shadowing performance at the critical rates must be the result, at least indirectly, of degraded recognition performance, which must in turn be the result, at least indirectly, of the particular rates of alternation. Thus, the degradation of the signal results in the subject's being *overloaded* and unable to perform both tasks. Similar "Archimedes" effects have been observed in other areas, for example in remembering strings of digits, where recall accuracy for words declines as perceptual load is increased (Dallett, 1964; Rabbitt, 1966). Savin and Perchonock (1965) even used an Archimedes task to measure the psychological complexity of sentences of varying syntactic complexity. There are also other examples in the literature.

Furthermore, the idea that the two tasks overload the subject when his performance on one of them falters offers an explanation for the sometimes large differences in level of performance of different subjects. One would expect subjects to differ in their total processing capacity.

Secondly, even if in some experiments a subject

could perceive the degraded speech, if only he did not have to repeat it, this is clearly false in some parts of the present experiments. In Figure 4, the shadowing scores represented by Curve A decline almost to zero. When a subject is repeating only 10 or 15 words correctly out of a 100-word passage, and the time pressure has been relieved because the inserted silences have slowed down the average speech rate, his shadowing performance can hardly be imposing a heavy load, because he simply is not saying very much. Presumably, his difficulties in shadowing the message must arise because he is unable to recognize it. If so, shadowing scores may be rather a good index of intelligibility. Further, since the perceptual load is now so heavy, it alone is sufficient to overload the subject, and the effect could be shown without using a difficult task. Similar results could probably be obtained for temporally segmented speech using short sentences, or even PB word lists. Wingfield and Wheale (Note 1) were able to use a procedure like this in their study of the effect of word rate on alternated speech. They spliced in a silent interval after every 10 words of the message, during which the subject repeated what he had just heard. The results they obtained are very similar to those obtained with shadowing, but one can argue that their task imposes a memory load not present in the shadowing task, and also that the pauses give the subject time to apply processing strategies that he has no time for in everyday speech. The trouble is, there is *no* way of studying the perception of running speech that produces quantifiable data and yet does not interfere with the normal processes of speech perception.

Effective Speech Rate

A second possible criticism of the present result is that the decline of intelligibility might follow from the progressive decrease in the effective rate of the speech, rather than the particular values of speech- and silent-interval duration. This possibility can be rejected by plotting the shadowing scores shown in Figure 4 as a function of the time taken to present 1 sec of original speech in its temporally segmented form. The speech-varying and silence-varying functions agree quite well up to a time multiplier of about four, but thereafter they diverge widely. This can also be inferred from a comparison, in Figure 4, of the left-hand side of Curve B with the right-hand side of Curve A, which are obviously bound for different targets, although their time factors are comparable.

Auditory Short-Term Storage

It is tempting to try to explain the results obtained with temporally segmented speech in terms of auditory short-term memory processes. Such processes have been invoked by several investigators (e.g., Békésy, 1971; Massaro, 1972; Neisser, 1967, among others), and called "echoic memory" or "pre-perceptual auditory storage."

The hypothesized echoic storage has two interesting properties. First, it operates at a very early stage in the signal's progression from ear to percept, and stores a representation of the signal that is relatively raw and unprocessed. This fits well with Siebert's (1968) observation that, in living information-processing systems, "it is obviously desirable to preserve as many as possible of the original stimulus details through at least the early stages of the data-processing system," since variations in the animal's immediate predicament may require that different aspects of the input signal should be discarded as inessential on different occasions. Storage of this type (if it exists) must precede any recognition of patterns, or properties, of the signal, since such pattern- or parameter-extraction is "fundamentally an information-destroying process," which discards any information in the stimulus that does not form part of the pattern. Secondly, the echoic store is of relatively short duration, and can perhaps be thought of as prolonging the otherwise fleeting existence of acoustic events.

Storage of this kind has obvious advantages for an organism bombarded with stimulation through all senses, but only a limited channel for detailed processing. It allows attention to be selective, and limits the need for detailed processing of nonattended inputs until something requiring attention occurs. In vision, events seen in peripheral vision usually have sufficient permanence that they are still available for fixation by the attracted foveal attention, so that raw storage may not be very important. But in audition, the sound that attracted attention has no permanence and raw storage becomes vital. With it, a brief acoustic event can both attract attention to itself and also be available (from the echoic store) for any detailed processing the attracted attention may decide is appropriate. This is quite similar to a popular solution to the problem of eliminating long silent intervals from tape recordings without clipping the front of the bursts of signal. The input drives a voice-operated relay which starts a tape recorder. The tape recorder receives its signal input through a delay longer than the recorder's start-up transient, so that the machine is ready to record the start of the input signal when it arrives over the path containing the delay. We will return to this analogy below.

A clear example of an experiment that bears directly on the existence and properties of the echoic store was reported by Guttman and Julesz (1963). They presented subjects with iterated segments of random noise, produced by a computer in such a way that the statistical properties of the signal were the same across the joins of adjacent iterations as within an iteration. When the iterated segment was longer than about 1 sec, repetitions could be detected only with effort, and detection grew progressively more difficult as the segment was lengthened. With iterated segments shorter than 1 sec, the perception of

repetition was always immediate and effortless. With segments shorter than 1 sec but longer than 250 msec, listeners described their percept as "whooshing." Since they were able to perceive detail *within* a whooshing segment, the memory system underlying the percept may be storing parameters derived from the raw signal rather than the raw signal itself.

With an iterated segment lasting less than about 250 msec but more than about 50 msec, subjects reported "motorboating," in which they could not discriminate detail within the iterated segment. Shorter segments gave rise to a sensation of pitch. [A finding that supports Guttman and Julesz's division of the continuum in this somewhat subjective way is Michon's (1964) measurement of JNDs for repetition rate of a train of pulses, which showed discontinuities in subjects' sensitivities at almost exactly the durations described by Guttman and Julesz.] Clearly, any memory system that can detect iterations up to 250 msec long in white noise must be storing a record of the signal that retains some of the microstructure of the noise—that is, it stores a relatively raw representation of the signal, and apparently stores it for at least 250 msec without appreciable decay.

Obviously, the reason for developing the foregoing argument is that it meshes very well with the results from the present experiments with temporally segmented speech, if one further assumes that all more-detailed processing looks to the echoic store for its input. Consider first the results with speech-varying temporally segmented speech, in which silent intervals lasting 200 msec were used to separate speech intervals of variable duration. If we assume that a silent interval this long fills the echoic store, then the acoustic events that precede and follow a silent interval are never simultaneously available in the echoic store. Therefore, the higher order processors are unable to compare them, but must treat the events within each speech interval independently of other intervals.

If it is necessary for acoustic events to coexist in the echoic store for higher order processors to be able to compare them, then it follows that no higher order processor *looking at the raw signal* can cope with events whose time span is too long to fit in echoic storage. Any higher order processors which *do* compare events over a longer time span must use derived descriptions of the signal as input, and not the raw signal. As long as speech-interval duration is longer than the capacity of echoic storage, any constraints on performance must be the result of the limited storage capacity. It is only when speech intervals become shorter than storage capacity that *they* become the limiting factor, and intelligibility begins to decline. Thus the speech interval duration at which intelligibility first begins to decline is a measure of the capacity of echoic storage. Inspection of Figure 4, Curve A, suggests a capacity of about 180 msec.

As speech interval duration is progressively shortened beyond this point, the higher order processors have less and less signal to look at, so it is not surprising that intelligibility progressively declines. It is not clear if anything can be concluded from the shape of the function that describes this decline of intelligibility. Its shape appears to suggest that each halving of speech-interval duration produces a constant decrement in intelligibility of about 30%, at least from 90% down to 10% intelligibility. This may reflect the distribution of speech cues as a function of their time span—and also, perhaps, the distribution of the time windows of the detectors that extract those speech cues. This argument suggests that the shape of the function may be specific to speech, in which case one might expect it to depend on the speech content, rather than the duration, of the speech intervals, since that was the result obtained with alternated speech (Huggins, 1964; Wingfield & Wheale, Note 1). A preliminary account of an experiment supporting this expectation appears elsewhere (Huggins, Note 2). On the other hand, some other results obtained with nonspeech signals agree remarkably well with the present result, which suggests that the shape of the function may reflect a more basic property of the auditory system. For example, Nixon, Raiford, and Schubert (1970) studied monaural phase perception in two-tone complexes by inserting, into a continuous tone pair in one phase relationship, a brief excerpt of a test pair in which the phase relation was either unchanged or inverted. Brief silent intervals separated the test tones from the background. The subject made a same/different judgment. The extent to which their subject's performance departed from chance agrees almost exactly with performance in the speech-varying task, when the results are compared as a function of the duration of the test-tone segment, or speech interval. Similarly, one of Massaro's subjects, in a study of backward recognition masking of tones (Massaro, 1972, Figure 1, Subject A.L.), gave results which, when corrected for chance performance, show a similar excellent agreement. (The same applies to his other subjects too, if their results are further corrected for an asymptotic performance less than 100%.)

Performance in each of these studies seems to depend on the duration of a sample of signal or intersignal silence. It is tempting to speculate that the reason the studies have similar results is that they tap a single property of the perceptual apparatus.

Gap-Bridging: A Self-Contained Phenomenon?

Now let us turn to the other half of the present result, the recovery of intelligibility as the silent intervals are shortened (Curve B in Figure 4). We will

consider two different interpretations for this result. Unfortunately, the data from the present experiment are not sufficient to decide between them.

The first interpretation assumes that the shape of the function describing the recovery, and its position along the abscissa, would not change if the duration of the *speech* intervals used to obtain it were changed. Of course, the asymptotic intelligibility, with long silent intervals, would depend on speech-interval duration, but whatever this level, intelligibility would start to recover when silent intervals were shortened below 120 msec, and the recovery would be complete when they reached 60 msec. Thus the gap-bridging is, under this interpretation, a self-contained phenomenon quite separate from the dependence of intelligibility on speech-interval duration. Other lines of research have suggested that there is something special about durations of around 60 msec in audition and elsewhere (Kristofferson, 1967; Stroud, 1955), and Efron (1970) has presented evidence that the minimum duration of a perception is about 120 msec. Clearly, further research is needed to discover whether the critical durations in the present experiment (63 and 125 msec of silence when speech-interval duration was 63 msec) are more than coincidentally related to the other effects mentioned, which used very different paradigms. A point in favor of the interpretation of gap-bridging as a self-contained phenomenon is the remarkable agreement between the predicted and observed shadowing scores shown in Figure 5. The prediction made use of values for P_B, the probability that the gap could be bridged, that were derived from Curve B in Figure 4. Thus an implicit assumption was made that gap-bridging depended only on silent-interval duration, and that varying the *speech*-interval duration would have no effect except on the asymptotic intelligibility obtained with long silent intervals.

On the other hand, interpreting gap-bridging as an independent phenomenon raises a problem. As mentioned above, increasing the speech rate of alternated speech moves the critical rates of alternation to proportionately higher values (i.e., shorter durations). I have argued above that the recovery of intelligibility at higher rates of alternation is due to gap-bridging—but if gap-bridging is an independent phenomenon, then the recovery of intelligibility should depend only on the duration of the silent intervals, and should not be affected by a change in speech rate. Reinspection of the figure which led to the conclusion that the intelligibility function was moved to higher rates of alternation by the speedup (Huggins, 1964, Figure 1) shows that although the data support the conclusion well at low rates of alternation, where we have argued that intelligibility depends on the speech intervals, the support is much less convincing at higher rates, where gap-bridging is dominant. On the other hand, Wingfield and Wheale's (Note 1) success in replicating the speed-up effect seems to hold across all the rates they studied. Thus, presently available data do not permit gap-bridging to be either accepted or rejected as an independent phenomenon.

Gap-Bridging: Related to Echoic Memory?

The second interpretation of gap-bridging relates it to the same echoic memory process that was invoked to interpret the dependence of intelligibility on speech-interval duration. Briefly, the argument is that silent intervals only disrupt intelligibility to the extent that they prevent higher order processors from comparing events before a silent interval with events after it. A silent interval of 200 msec was sufficient to fill echoic memory, and intelligibility began to decline when speech interval duration was reduced to below 180 msec. When speech intervals are shortened to 63 msec, intelligibility falls to 55%. If *silent* intervals are now shortened, intelligibility is unaffected until the silences last 125 msec. At this point, echoic memory, with a capacity of 180-200 msec, has room for one speech interval of 63 msec and one silent interval of 125 msec. If the silent intervals are further shortened, there now begins to be room for a second speech interval, following the silent interval. Gap-bridging has begun, and intelligibility improves. When silent intervals reach 63 msec, there is room in echoic storage for a complete speech interval, plus a complete silent interval, plus a second complete speech interval. Gap-bridging is complete, and intelligibility is completely recovered.

Although the foregoing description is appealing, it is also incomplete. For example, it does not explain why no increase in intelligibility occurs with silent intervals of 125 msec, although this should leave room in echoic storage for the second *half* of the earlier speech interval, and the first *half* of the later one. Perhaps some synchronizing effect should be invoked.

The interpretation just presented has two strong points in its favor. First, it is clearly testable, and experiments are in fact under way to test it. Second, there is also data to support it from a recent experiment using dichotically alternated clicks (Huggins, 1974). Subjects adjusted the repetition rate of a diotic pulse train so that its perceived rate matched the perceived rate of a dichotically alternated train. At low rates of alternation, subjects matched the total rate into the head of the dichotic train, and each pulse in the alternated train was perceived as a separate event, delimiting an *interval*. At high rates, subjects matched the *rate* in one ear of the alternated train, and perceived a pulse train at each ear. The argument can be made that the interval mode of perception obtains when the interval between successive pulses in one ear is long enough that two pulses *never* coexist in echoic memory for that ear. (Note that separate echoic storage for the two ears was

one of the implications of the earlier work with alternated speech, cf. Huggins, 1964.) When interpulse interval is short enough that two pulses *always* coexist in echoic memory, the rate mode of perception becomes stable. Thus the interval mode should start to break down when a new pulse enters echoic storage just as the earlier pulse leaves it—that is, when interpulse interval *in one ear* is equal to the capacity of echoic storage. Further, the rate mode should become stable when a new pulse enters echoic storage just as the pulse *two* before leaves it—that is, when the interpulse interval in one ear is equal to half the capacity of echoic storage. This leads to two predictions: the interval mode should start to break down at a pulse separation of about 200 msec in one ear (i.e., 100 msec in the alternated train), and the crossover between interval and rate modes should start and end at pulse intervals that stand in a radio of 2:1. The data for two of the three subjects are in good agreement with both predictions, and the data for the third subject, which is more noisy, are in fair agreement (Huggins, 1974, Figure 1).

This interpretation of gap-bridging also provides a neat resolution of what would otherwise be a conflict with Massaro's (1972) results on backward recognition masking of tones. In Massaro's experiment, a 20-msec sample of tone was followed after a variable silent interval by 500 msec of a masking tone, and the subjects' task was to identify the first burst as being (6%) higher or lower in frequency than the masker. Subjects' performance *improved* as the silent interval was lengthened. But in the gap-bridging experiment, subjects' ability to identify the speech in a 60-msec sample *declined* as the silent interval following it and preceding the next speech interval was lengthened. Why were the contents of the first interval not masked by the occurrence of the second, when the silent interval separating them was short? The answer suggested by the arguments developed above is that, in both cases, the ear tries to integrate into a single percept any two relatively similar events that coexist in echoic storage, and only becomes able to treat them as separate events if they do *not* coexist in echoic storage. Thus, when the silent interval is short, (1) successive speech samples are integrated into a single event, and intelligibility is high, and (2) target and masker are integrated into a single event, and the subject is unable to treat the target as a separate percept and identify it. The reverse is true when the silent intervals are long. This also explains why, in speech, a vowel does not mask the consonant that precedes it, as one would expect from the results in backward recognition masking. It may also cast light on recent studies in the perception of temporal order (Warren, Obusek, & Farmer, 1969) and auditory stream segregation (Bregman & Campbell, 1971).

Some Final Speculations

The studies that established the properties of primary visual storage, or "iconic" storage (Neisser, 1967), showed that the strength of the stimulus trace, or at least its availability, decays quite rapidly in the store. On the other hand, the results and interpretations presented above suggest a model for echoic storage with rather different properties. To be explicit, echoic storage seems to have properties more like those of a delay line, in which information entered at the front of the line travels down it without degradation, until it reaches the end, where it is lost. Such a difference in storage becomes quite reasonable if one considers the differences between vision and audition. In vision, the input is a spatial pattern, and changes in this pattern in time are analyzed only rather coarsely, as evidenced by flicker-fusion effects and by the success of the movie industry. In audition, the input is a pattern of pressure-changes over time. Since time is inextricably involved in the specification of the stimulus before it is even transduced, one can argue that the time pattern simply could not be adequately stored in a system that allowed changes to occur over time, since this would result in changes in *quality* rather than quantity.

Earlier in the discussion, I argued that a selective attention mechanism must involve some peripheral storage if an event is to attract the attention of the selector and still be available for more detailed processing after attention is selected. An analogy was drawn with a tape recorder operated by a voice-controlled relay, but receiving its to-be-recorded signal over a path containing a delay. For either of these systems to work, there has to be a second transmission channel, with different properties. It might be referred to as a control channel, since it is responsible for starting the tape recorder, or for preparing the higher order processors for the arrival of information over the signal path containing the delay. Obviously the control information has to arrive before the signal, so transmission must be faster over the control channel—or at least it must not be delayed. The information transmitted over the control channel can be relatively coarse, since the higher order processors that do the detailed analysis use the undegraded signal arriving over the signal channel for their input. Gross details of the input are probably preserved in the control channel, since this information will be needed to guide the decision of what detailed processing of the input is appropriate. The information in the control channel may not normally be available to consciousness, since this would reduce the advantage gained by having a selective attention. However, the control information might be available to the extent that it could trigger a preprogrammed response, of the sort required in an experiment on simple reaction time, for example.

In developing this tentative account of early stages in auditory processing, I have tried to parallel as closely as possible a model described by Wall (1970) for the function of the dorsal columns of the spinal cord. Transmission of sensory information from the periphery to the brain seems to take place over two separate paths, one fast and one slow. Wall's experiments, and many observations that he reports, suggest that the purpose of the fast path, whose contents do not normally reach consciousness, is to set up the "analysis programs" for the more detailed neural representation arriving over the slow pathway. Wall's description was developed to account for a body of data that earlier theories of dorsal-column function could not explain. His arguments are tightly reasoned, and are supported by a wide range of experiments and observations.

The attempt to extend Wall's model to audition was not motivated by a body of intractable data, but rather was undertaken because it provides an appealing functional framework that seems to tie together several different lines of research. Several of these have been described briefly above, and the possibility of a second signal channel in audition has been suggested on other grounds by Bernstein (1970), among others.

REFERENCE NOTES

1. Wingfield, A., & Wheale, J. L. *Word-rate and intelligibility of alternated speech*. Manuscript submitted for publication, 1975.

2. Huggins, A. W. F. *More temporally segmented speech: Is duration or speech content the critical variable in its loss of intelligibility?* Research Laboratory of Electronics, Massachusetts Institute of Technology, Quarterly Progress Report, 1974, 114, 185-193.

REFERENCES

Békésy, G. von. Auditory backward inhibition in concert halls. *Science*, 1971, **171**, 529-536.

Bernstein, I. H. Can we see and hear at the same time? In A. F. Sanders (Ed.), Attention and performance III. *Acta Psychologica*, 1970, **33**, 21-35.

Bregman, A. S., & Campbell, J. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, 1971, **89**, 244-249.

Cherry, E. C., & Taylor, W. K. Some further experiments upon the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 1954, **26**, 554-559.

Dallett, K. M. Intelligibility and short-term memory in the repetition of digit strings. *Journal of Speech and Hearing Research*, 1964, **7**, 362-368.

Efron, R. The relationship between the duration of a stimulus and the duration of a perception. *Neuropsychologica*, 1970, **8**, 37-55.

Guttman, N., & Julesz, B. Lower limits of auditory periodicity analysis. *Journal of the Acoustical Society of America*, 1963, **35**, 610(L).

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.

Huggins, A. W. F. The perception of temporally segmented speech. In *Proceedings of the VII International Congress on Phonetic Sciences. Montreal.* The Hague: Mouton, 1972. Pp. 531-535.

Huggins, A. W. F. On perceptual integration of dichotically alternated pulse trains. *Journal of the Acoustical Society of America*, 1974, **56**, 939-943.

Kristofferson, A. B. Attention and psychophysical time. In A. F. Sanders (Ed.), Attention and performance. *Acta Psychologica*, 1967, **27**, 93-100.

Massaro, D. W. Preperceptual images, processing time, and perceptual units in auditory perception. *Psychological Review*, 1972, **79**, 124-145.

Michon, J. A. Studies on subjective duration I: Differential sensitivity in the perception of repeated temporal intervals. *Acta Psychologica*, 1964, **22**, 441-450.

Neisser, U. *Cognitive psychology*, New York: Appleton-Century-Crofts, 1967.

Nixon, J. C., Raiford, C. A., & Schubert, E. D. Technique for investigating monaural phase effects. *Journal of the Acoustical Society of America*, 1970, **48**, 554-556.

Rabbitt, P. Recognition: Memory for words correctly heard in noise. *Psychonomic Science*, 1966, **6**, 383-384.

Savin, H. B., & Perchonock, E. Grammatical structure and the immediate recall of English sentences. *Journal of Verbal Learning and Verbal Behavior*, 1965, **4**, 348-353.

Siebert, W. M. Stimulus transformations in the peripheral auditory system. In P. Kolers & M. Eden (Eds.), *Recognizing patterns*. Cambridge: M.I.T. Press, 1968. Pp. 104-133.

Speaks, C., & Trooien, T. T. Interaural alternation and speech intelligibility. *Journal of the Acoustical Society of America*, 1974, **56**, 640-644.

Stroud, J. The fine structure of psychological time. In H. Quastler (Ed.), *Information theory in psychology*, New York: Free Press, 1955.

Wall, P. D. The sensory and motor role of impulses travelling in the dorsal columns towards cerebral cortex. *Brain*, 1970, **93**, 505-524.

Warren, R. M., Obusek, C. J., Farmer, R. M., & Warren, R. P. Auditory sequence: Confusions of patterns other than speech or music. *Science*, 1969, **164**, 586-587.

(Received for publication March 5, 1975;
revision received April 20, 1975.)