Journal of Experimental Psychology: Human Perception and Performance
1975, Vol. 104, No. 2, 121–129

Perception of Temporal Order in Vowel Sequences With and Without Formant Transitions

Michael F. Dorman, James E. Cutting, and Lawrence J. Raphael
Haskins Laboratories, New Haven, Connecticut

Temporal-order perception of phoneme segments in running speech is much superior to temporal-order perception in repeating vowel sequences. The more rapid rates possible in running speech may be due largely to the presence of formant transitions. In a series of five experiments we observed that many temporal-order misjudgments of repeating vowels can be explained in terms of auditory stream segregation, triggered for the most part by discontinuities in first-formant frequencies of adjacent vowels. Streaming, however, can be suppressed by formant transitions appropriate for the perception of stop consonants and by continuous transitions resembling those in coarticulated vowels. At rapid sequence rates, when the constraints of auditory streaming are removed, correct temporal-order identification is limited by linguistic transformations of vowels into other phoneme segments.

Two distinguishing characteristics of speech perception are the rate at which speech can be perceived (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967) and the proficiency with which temporal-order information is preserved. The present article is concerned primarily with the second of these characteristics. We shall attempt to explain the difference between the relatively poor temporal-order performances reported in the literature for certain repeating sequences and the considerably better temporal-order performance for running speech.

If listeners are to identify the temporal order of a repeating sequence of four concatenated steady state vowels, each vowel must be 125 to 250 msec long (Thomas, Hill, Carrol, & Garcia, 1970; Warren & Warren, 1970). Since 30-msec vowels can be accurately identified in isolation, Thomas et al. (1970) concluded that the duration of each vowel segment in excess of 30 msec is used for encoding or decision making. This view and its supporting data fit nicely into a theory of speech perception which suggests that perceptual processing of vowels can last between 120 and 250 msec (Massaro, 1972).

The perception of concatenated speech segments, however, may differ greatly from the perception of speech segments in running speech. For example, listeners can comprehend speech at a rate of up to 400 words/min, or approximately 30 phonemes/sec, without temporal-order confusions (Orr, Friedman, & Williams, 1965). This rate translates into an average phoneme duration of only 30 to 40 msec, an estimate markedly different from that derived from Warren-type repeating vowel sequences. The discrepancy between the two estimates of the minimum phoneme duration that permit accurate temporal-order judgments prompted a series of experiments with repeating speech sequences.

One factor contributing to the difference between temporal perception abilities in certain repeating vowel sequences and in running speech would appear to be formant transitions: The stimuli of Thomas et al. (1970) and Warren and Warren

This research was supported by Grant HD-01994 to the Haskins Laboratories from the National Institute of Child Health and Human Development. Michael F. Dorman and Lawrence J. Raphael are also at Herbert Lehman College of the City University of New York. James E. Cutting is also at Wesleyan University, Middletown, Connecticut.

Requests for reprints should be sent to the authors at the Haskins Laboratories, 270 Crown Street, New Haven, Connecticut 06510.

TABLE 1

Formant Frequency Values (in Hz) for the Four Vowels Used in All Experiments

Vowel		Formant	
	First	Second	Third
/i/ (EE as in *heed*)	386	2,234	2,862
/æ/ (AA as in *had*)	666	1,695	2,525
/ɔ/ (AW as in *hawed*)	614	846	2,348
/u/ (UU as in *who'd*)	386	769	2,180

(1970) were physically abutted with considerable discontinuities between the formants of adjacent vowels, whereas in running speech few such discontinuities occur, in part because of articulatory constraints. Cole and Scott (1973) found that formant transitions between fricative and vowel segments in fricative–vowel syllables aided in the perception of temporal order when such syllables were placed in repeating sequences. These transitions, however, were not necessary for the identification of the syllables when presented in isolation. Thus, Cole and Scott concluded that formant transitions serve to integrate the speech stream beyond their role as carriers of phonetic information. We concur with this view and have attempted to extend it beyond fricative–vowel syllables, whose periodic and aperiodic components can easily be detached perceptually from one another, to syllables that consist entirely of formant resonances.

Experiment 1

Method

Warren-type repeating sequences were generated on the Haskins Laboratories' parallel-resonance synthesizer. The sequences consisted of long steady state vowel syllables /i, æ, ɔ, u/ and of consonant-vowel-consonant (CVC) syllables /bib, bæb, bɔb, bub/. All stimuli were 120 msec long, well below the critical duration (168 msec) noted by Thomas et al. (1970) for 75% correct identification of repeating synthetic vowels. All stimuli had the same fundamental frequency (110 Hz) and overall amplitude contour (10 msec rise and fall). Steady state frequencies for vowels are shown in Table 1. Initial and final formant transitions in the CVC syllables were 45 msec long, leaving 30 msec of steady state vowel. Typical vowel and CVC sequences are shown at the left-hand side of Figure 1. Stimuli within the same class were permuted in the six possible sequence orders. These 12 continuous, repeating sequences were recorded on audio tape with 10 msec between successive syllables, thus yielding a phoneme rate of about 8/sec for vowels and about 16/sec for CVCs. It should be noted that closure durations between stop consonants are typically much greater in running speech, and that with no final consonant release the syllables /bib, bæb, bɔb, bub/ sound like /bibæbɔbub/. Each sequence began at a very low volume, gradually increased in volume over the course of 5 sec to a maximum intensity (approximately 80 db. re $20 \mu N/m^2$), remained at that maximum for 10 sec, and then decreased to its original low volume during the final 5-sec period. This procedure eliminates the use of first and last syllables as anchors for determining sequence order. The tape was played twice for all listeners.

Twenty-two Yale undergraduate students participated in the task as part of a course requirement. The stimuli were reproduced on an Ampex AG500 tape recorder via an Ampex 620 loudspeaker. Tokens of the steady state vowels at 2-sec durations were played to the listeners until they could accurately identify the vowels. The listeners were then told that they would hear more rapid vowel and CVC sequences, and were instructed to write as soon as possible the identity of the vowels in the order that they heard them (disregarding the /b/s in the CVC stimuli).

Results and Discussion

Table 2 displays the average performance of the listeners for long-vowel and CVC stimuli for each of the six sequence orders. The vowel /i/ is arbitrarily designated as the first vowel of each sequence, but listeners could respond beginning with any vowel. In terms of performance summed over all sequence orders, the two classes of stimuli did not differ significantly: long-vowel and CVC sequences were correctly identified on 67% and 70% of all trials, respectively. This result, however, is somewhat deceptive, since it

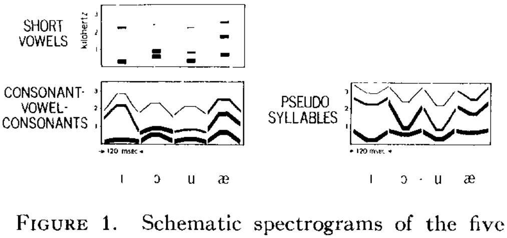

Figure 1. Schematic spectrograms of the five types of stimulus sequences used in the five experiments, in Sequence Order 4.

sums over sequence orders with quite varied performance levels.

Two orders of particular interest, Order 2 /i, æ, u, ɔ/ and Order 4 /i, ɔ, u, æ/, were more difficult than others. (Sequence Order 5 was also difficult to identify, and we will return to it in Experiments 2 and 4.) Taken together Sequence Orders 2 and 4 yielded a most interesting pattern: Long-vowel sequences at about 8 phonemes/sec were identified on only 51% of such trials, whereas CVC sequences at 16 phonemes/sec were identified on 64% of such trials, a result exactly opposite to what might have been predicted on the basis of the number of phonemes to be identified per unit of time. This pattern occurred against a background of small differences between the two stimulus classes for the other four sequence orders: 75% for vowel stimuli and 73% for CVCs. Neither of these differences, however, reached statistical significance.

Listeners readily volunteered that Orders 2 and 4 were difficult, and described the difficulty in terms of the sequences "flying apart." A typical report was that two groups of vowels were heard: one repeating group was /i/ and /u/ and the other was /æ/ and /ɔ/, with no apparent interconnection between the groups. We recognized this subjective quality as the hallmark of auditory stream segregation.

Bregman and Campbell (1971) reported that when listeners are presented a repeating sequence of six brief (100-msec) tones which alternate between high and low frequencies, they are unable to report correctly the high–low sequence. Instead they report two "streams" of tones, one containing the high-frequency items and the other containing the low-frequency items. Within a stream temporal identification is reasonably accurate (73%–79%). Between-stream identification, however, is no better than chance. Bregman and Campbell termed this phenomenon "primary auditory stream segregation." The perceptual experience of listening to Vowel Sequences 2 and 4 in the present study appears very similar in nature; although Sequences 2 and 4 are characterized by physically separated /i/–/u/ and /æ/–/ɔ/ items, listeners reported hearing each as a unit pair. Our phenomenon appears to be explainable in terms of perceptual streaming of first formants, the most prominent and lowest frequency component of the four vowels. As shown in Table 1, the first-formant frequency value for both /i/ and /u/ was 386 Hz, whereas for /æ/ and /ɔ/ it was 666 Hz and 614 Hz, respectively. Since /i/ and /u/ share low-frequency first formants and /æ/ and /ɔ/ have considerably higher frequency first formants, the vowels could form separate streams in a rapid sequence like that shown at the top left-hand side of Figure 1. Only Sequence Orders 2 and 4 meet the requirements of having alternating high and low first formants, and these are the orders that were most difficult for listeners to identify.

TABLE 2

Mean Percentage of Identification of Sequence Orders for Experiment 1

	Stimulus class		
Sequence order	Long vowels	Consonant-vowel-consonants	*M*
1. /i·æ·ɔ·u/	86	80	83
2. /i·æ·u·ɔ/	45	57	51
3. /i·ɔ·æ·u/	84	69	76
4. /i·ɔ·u·æ/	57	70	63
5. /i·u·æ·ɔ/	60	67	64
6. /i·u·ɔ·æ/	70	77	74
M	67	70	

A second important observation is that the temporal order of 30-msec vowels in the context of initial and final /b/ could be identified as accurately as the 120-msec vowels. Since 30 msec is far below the vowel duration necessary for accurate temporal identification of concatenated vowels (Thomas et al., 1970), the formant transitions in the CVC sequence appear to facilitate identification of temporal sequence. In this sense formant transitions and silent intervals may act in a similar manner. Warren and Warren (1970) reported that, although sequences of four concatenated 200-msec vowels are very

TABLE 3
MEAN PERCENTAGE OF IDENTIFICATION OF SEQUENCE ORDERS FOR EXPERIMENT 2

	Stimulus class			
Sequence order	Long vowels	Short vowels	Consonant-vowel-consonants	*M*
1. /i·æ·ɔ·u/	88	50	88	75
2. /i·æ·u·ɔ/[a]	44	38	78	53
3. /i·ɔ·æ·u/	50	13	50	38
4. /i·ɔ·u·æ/[a]	39	50	67	52
5. /i·u·æ·ɔ/	63	75	88	75
6. /i·u·ɔ·æ/	50	63	50	54
M	52	47	71	

[a] Represented twice as often as other sequence orders.

difficult to identify, the same sequences with 150-msec vowels separated by 50 msec of silence are relatively easy to identify.

Since the results of the present study suggest, but do not confirm, that certain sequence orders are more difficult to identify than others, Experiment 2 was designed, in part, to observe such differences in greater detail. In addition, Experiment 2 was designed to compare the relative contributions of transitions and silence in temporal-order identification.

EXPERIMENT 2

Method

The long-vowel and CVC stimuli from Experiment 1 were used again. In addition, short steady state vowel stimuli were synthesized. They were identical to the long vowels in all respects except duration. Whereas long vowels were 120-msec in duration, short vowels were only 30-msec long, and thus identical to the steady state portion of the CVCs. The other 90 msec of the stimuli was replaced by silence. Preliminary tests revealed that the stimuli in all three classes were at least 90% identifiable in isolation. Again, stimuli within a class were permuted in the six possible sequence orders, but Orders 2 and 4 were represented twice as often as the other four. Schematic spectrograms of the long-vowel, short-vowel, and CVC stimuli in the order /i, ɔ, u, æ/ are shown at the left-hand side of Figure 1. Relative amplitudes of each formant are indicated in terms of formant width, and formants are numbered ordinally from low frequency to high. Twenty-four sequences were recorded in the same fashion as in Experiment 1, with class of stimuli and sequence order randomly intermixed.

Eight students at Herbert Lehman College of the City University of New York and three staff members from the Haskins Laboratories (not including the authors) served as listeners. Stimuli were reproduced for the Lehman College listeners on a Revox 1122 tape recorder via AR-4x loudspeakers, and for the Haskins listeners on the same apparatus as in Experiment 1.

Results and Discussion

On the average the long-vowel and short-vowel sequences were considerably more difficult to identify than the CVC sequences, as shown in Table 3: Respective performance levels were 52%, 47%, and 71%. CVC sequences were significantly easier to identify than long-vowel sequences, $T(8) = 3$, $p < .05$, and short-vowel sequences, $T(8) = 5$, $p < .05$. Such differences are even more striking in Sequence Orders 2 and 4: Taken together, long-vowel sequences were correctly identified on only 42% of all these trials and short-vowel sequences on 44%, whereas CVC consonants were identified correctly on 77%. These group averages accurately reflect individual performances; for example, six listeners achieved perfect performance on these two CVC sequences, whereas only three achieved perfect performance on the corresponding long-vowel sequences and three on the short-vowel sequences. Perhaps the nonsignificant difference between the CVCs and long-vowel sequences in Experiment 1 is attributable to a ceiling effect induced by over-representation of the easier sequences 1, 3, 5, and 6.

We have no explanation for the apparent extreme difficulty of short-vowel Sequence Order 3. Note, however, that Sequence Order 5, which had been as difficult as Order 4 in Experiment 1, was one of the easier sequence orders to identify in this experiment. There were no systematic differences between Lehman and Haskins listeners.

There is an apparent discrepancy between our results and those of Warren and Warren (1970). We found that long-vowel and short-vowel sequences were equally difficult to identify, whereas Warren and Warren, using 200-msec concatenated vowels and the same vowels with 50 msec replaced by silence, found the latter to be easier to identify in sequence. Perhaps the ratio of vowel duration to silence duration is important here. In the present study the vowel:silence ratio was 1:3, whereas in the Warren's study it was 3:1.

The superiority of the CVC sequences over the short-vowel sequences indicates that transitions between vowel nuclei are more effective than silence in reducing auditory stream segregation. This outcome is encouraging, since few such silent intervals occur in running speech, yet correct temporal identification of phoneme order is effortlessly achieved by the listener.

Perhaps there are ways of perceptually "gluing" vowels together with formant transitions other than those appropriate for stop consonants. Experiment 3 was designed, in part, to determine whether or not streaming could be overcome through the use of longer, continuous transitions between vowel nuclei. This general tactic has proved useful in limiting primary auditory stream segregation of pure tones (Bregman & Dannenbring, 1973) and in fricative–vowel syllables (Cole & Scott, 1973). Experiment 3 was also designed to determine if streaming is suppressed by all transitions or only by transitions that are phonetically and articulatorily reasonable.

Experiment 3

Method

The long-vowel and CVC stimuli were used again. In addition, two other sets of stimuli were generated. Both sets contained 30-msec steady state vowel nuclei corresponding to the vowels /i, æ, ɔ, u/, and both had initial and final formant transitions. In one set the transitions were context dependent, gliding continuously (except for a 5-msec break) over the course of 95 msec from the steady state formant values of one vowel into the succeeding vowel. These sequences are termed connected-vowel sequences. The other set consisted of pseudosyllables, which contained most of the features of the CVC stimuli except that the formant transitions were turned "upside-down"; that is, instead of all transitions gliding upwards in frequency into the vowel and downwards after it (appropriate for the perception of /b/), all transitions glided downwards into the vowel and back upwards after it (inappropriate for the perception of any consonant phoneme). Only three of the six possible sequence orders were selected: Orders 2 and 4 to maximize error probability, and Order 1 for comparison purposes. Schematic versions of the four sequence classes in the vowel order /i, ɔ, u, æ/ are shown in the four corners of Figure 1. Class of stimuli and sequence order were randomized and recorded as repeating sequences on audio tape. Each possible sequence occurred twice, yielding 24 sequences. The listeners were nine Lehman College students. Stimuli were reproduced on a Revox 1122 tape recorder via an AR-4x loudspeaker. In all respects the procedure was the same as in the two previous studies.

TABLE 4

Mean Percentage of Identification of Sequence Orders for Experiment 3

	Stimulus class				
Sequence order	Long vowels	Connected vowels	Consonant-vowel-consonants	Pseudosyllables	*M*
1. /i·æ·ɔ·u/	50	55	44	11	40
2. /i·æ·u·ɔ/	6	78	83	17	46
4. /i·ɔ·u·æ/	33	65	55	11	41
M	30	65	61	13	

Results and Discussion

As shown in Table 4, there was a large difference between the accuracy of sequence identification for the two types of vowel stimuli. Long-vowel sequences were identified correctly on only 37% of all presentations, whereas connected-vowel sequences were identified on 65%. All subjects demonstrated this trend ($T(9) = 0$, $p < .01$). The CVCs were also identified more accurately than the long vowels, replicating the results of Experiment 2, and again all listeners demonstrated this effect ($T(9) = 0$, $p < .01$). There was no difference between CVC and connected-vowel sequence identifications. The pseudosyllable sequences were essentially incomprehensible, and were identified at chance performance level. This appears to be primarily due to the fact that the vowels of isolated pseudosyllable items are extremely difficult to identify.

Interpolating gradual formant transitions between the 30-msec vowel nuclei was successful in inhibiting the streaming effect, a result which parallels that of Bregman and Dannenbring (1973). These authors reduced streaming in nonspeech signals by connecting steady state frequencies with smooth frequency ramps.

TABLE 5

Mean Percentage of Identification of Sequence Orders for Experiment 4

	Formants with transitions								
	Long vowels	Intermediate sequence types						Connected vowels	*M*
Sequence order	0	1	2	3	1–2	1–3	2–3	1–2–3	
2. /i·æ·u·ɔ/	29	71	71	63	96	79	71	91	71
4. /i·ɔ·u·æ/	54	87	75	46	88	84	79	88	75
5. /i·u·æ·ɔ/	67	63	71	67	67	67	71	83	70
M	50	74	72	59	84	77	74	87	
			($\bar{X}$ = 68)			($\bar{X}$ = 78)			

However, the speech data of the present experiments and the nonspeech data of Bregman and Dannenbring differ in several interesting ways. First, when listening to nonspeech signals, listeners did not achieve perfect performance on *same–different judgments* of two different ramped-tone sequences even when the steady state duration of the tones was 225 msec. In the present study, however, six of nine listeners achieved perfect performance or made only one error in *absolute identification* of connected-vowel sequences when steady state vowel duration was only 30 msec. Second, in the present study CVC sequences reduced streaming as well as the connected-vowel sequences, yet in CVCs transitions are not linear from vowel to vowel. It remains to be demonstrated whether or not such nonlinear ramps will reduce nonspeech auditory streaming. Since identification of the pseudosyllable stimuli proved very difficult, the hypothesis that only phonetically relevant transitions reduce auditory streaming of speech sequences remains unconfirmed.

We have demonstrated, like Bregman and Dannenbring (1973) and Cole and Scott (1973), that streaming can be inhibited through the use of transitions. We have argued that streaming of vowel sequences is caused primarily by discontinuities in first-formant frequencies of adjacent vowels. This hypothesis, however, has yet to be empirically substantiated. Therefore, in Experiment 4 we explored the relative contributions of the first, second, and third formants to the auditory streaming of vowel sequences.

Experiment 4

Method

Three sequence orders were selected: Orders 2 and 4 with interleaved /i/–/u/ and /æ/–/ɔ/ pairs, and Order 5, which had previously given some difficulty in temporal-order identification (Experiment 1). Each sequence was generated in eight different renditions: long-vowel sequences with no connecting transitions, connected-vowel sequences with transitions between all three formants of the vowel nuclei, and six other intermediate sequence types. Three sequence classes had just one formant connected, Formants 1, 2, or 3; three others had two formants connected, Formants 1 and 2, 1 and 3, or 2 and 3. The three different sequence orders of the eight different types of sequence classes were randomly intermixed and recorded in the same fashion as in previous experiments. Stimuli were played over the same apparatus as in Experiment 1. Twelve members of the staff of the Haskins Laboratories listened via an Ampex 620 loudspeaker to two passes through the tape, yielding 48 sequence identifications per subject.

Results and Discussion

The matrix of results is shown in Table 5. Consider first the long-vowel and the connected-vowel stimulus sequences. For the two sequence orders of particular interest, 2 and 4, the identification of the connected-vowel sequence is clearly better than that for the long-vowel sequence, an average of 41% and 89%, respectively. The differences were significant for both Orders 2 and 4, $T(10) = 0$, $p < .01$, and $T(9) = 0$, $p < .01$, respectively. No significant difference occurred for Sequence Order 5, although the result is in the same direction.

Of the sequences with only one formant connected, those with third-formant transitions were more difficult to identify than those with first or second formants connected. There was, however, no significant difference between contributions of first- and second-formant transitions. For Sequence Order 4 the results go in our predicted direction (Formant 1 suppressed streaming better than Formant 2), but for Sequence 2 there was no difference. Moreover, for Sequence Order 5 second-formant transitions alone suppressed streaming better than first-formant transitions. The reason for this may lie in the particular organization of formant resonances in this sequence order. Consider the long-vowel sequence /i, u, æ, ɔ/ in conjunction with the formant frequencies shown in Table 1. Notice that the first formants do not alternate between high and low frequencies, but that the second formants do. Perhaps for this sequence order it is second-formant discontinuities that facilitate streaming. None of the observed differences, however, are significant.

Consider next the sequences in which two formants were connected with transitions. Overall performance here was better than for the one-formant sequences, but not as good as that for the connected-vowel sequences. Third-formant contributions appear to be minimal. Compare (a) sequences in which only the first formant is connected with those in which first and third formants are connected, (b) second-formant-connected sequences with second- and third-formant sequences, and (c) first- and second-formant-connected sequences with connected-vowel sequences. Each comparison reveals no more than three percentage points advantage when adding the third-formant transitions, and none of these differences is significant. Since the third formant appears to contribute little to the inhibition of auditory streaming, the comparison of sequences with first and third formants connected and those with second and third formants connected may shed light on the comparative contributions of the first two formants to the streaming phenomenon. Again, there is no significant difference between the two sequence types. The pattern here, however, is nearly identical to the one-formant-connected sequences. For the two sequence orders of general interest, Orders 2 and 4, the first- and third-formant-connected sequences were somewhat easier to identify than the second- and third-formant-connected sequences. Again, Sequence Order 5 reverses this trend, suggesting that for this particular sequence the linkage of second formants is more important than that of first formants.

Experiment 5

Since connected-vowel sequences resist streaming significantly better than long-vowel sequences, in our final experiment we sought to assess the magnitude of the effect and to determine the minimum stimulus duration in connected-vowel sequences that permits the identification of temporal order.

Method

The basic stimuli were long-vowel and connected-vowel items in Sequence Orders 2, /i, æ, u, ɔ/, and 4, /i, ɔ, u, æ/. For both vowel classes and both sequence orders repeating sequences were synthesized, varying in duration from 1440 to 280 msec per four-item sequence in 13 equal steps. For long-vowel sequences individual vowel durations ranged from 360 to 72 msec, and for connected-vowel sequences the vowel-nucleus duration ranged from 90 to 18 msec with corresponding transitions from 270 to 54 msec. The stimuli were blocked for presentation in terms of stimulus type (long vowel or connected vowel), sequence order (2 or 4), and presentation of sequence (ascending series—from 288- to 1440-msec duration—or descending series). These conditions were counterbalanced within and across listeners. Sequences were played to eight staff members of the Haskins Laboratories by on-line interaction with the computer-driven speech synthesizer. Stimuli were presented via Telephonics earphones (Model THD-39).

All listeners were pretrained. First a verbal description of the streaming phenomenon was given. Then subjects were told the correct order of a vowel sequence and were presented repeating sequences of the long-vowel stimuli beginning with the longest sequence and progressing incrementally to the shortest sequence. Subjects were instructed to notice when the order of the vowels became difficult or impossible to identify correctly. Practice with a connected-vowel sequence was given in the same manner. After several demonstrations all subjects reported that they understood the general

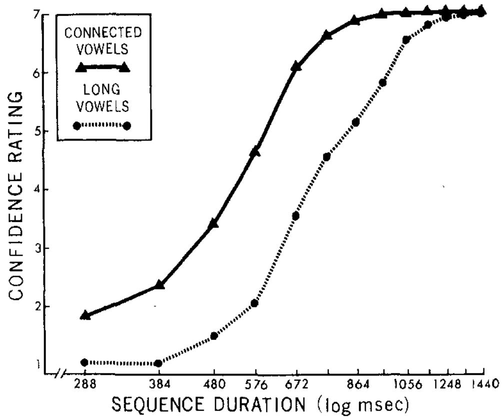

FIGURE 2. Mean of median confidence ratings as a function of sequence duration for long vowels and connected vowels. (A confidence rating of 1 indicates streaming. A confidence rating of 7 indicates that the sequence does not stream.)

phenomenon and recognized streaming sequences. They were instructed to report for each sequence in the ascending and descending series (in a method similar to that of Neisser & Hirst, 1974) whether the vowels streamed or were difficult to identify for any reason. They used confidence ratings for this judgment from 1 to 7 (1 indicated complete streaming and 7 no streaming).

Results and Discussion

Shown in Figure 2 are the mean of median confidence ratings for long-vowel and connected-vowel sequences as a function of sequence duration. As in Experiments 3 and 4, connected-vowel sequences resisted streaming significantly better than long vowels, $T(8) = 0$, $p < .01$, for the summed mean ratings at 13 sequence durations. The difference between the curves is indicative of the rating pattern for all subjects. If a rating of 4 is considered the threshold for streaming, connected-vowel sequences stream at durations of about 500 msec, whereas long-vowel sequences stream at about 700 msec. There was no difference between Sequence Orders 2 and 4.

Although both long-vowel and connected-vowel sequences were difficult to identify at brief durations, most subjects reported different problems in identifying temporal order. For the long-vowel sequences either the four vowels perceptually disordered themselves as in the previous experiments or, at the shortest durations, the vowels began to merge into a periodic buzz. On the other hand, as the connected-vowel sequences were made shorter, many verbal transformations were reported: /i, æ, u, ɔ/ became /yæwɔ/, and even /u, æ, i, ɔ/ became /pætio/ (*patio*) or /redio/ (*radio*). Here, identification of temporal order became difficult because one or more of the vowels seemed to disappear or change into another segment. In the first example cited, the failure to report /i/ or /u/ does not appear to be the result of backward masking, as Massaro (1972) might suggest, but rather may be seen as an instance of acoustic cues becoming appropriate for the perception of new segments: A brief /i/ gliding into /æ/ becomes /yæ/, and similarly a brief /u/ gliding into /ɔ/ becomes /wɔ/. For other transformations the rules may be similar. Thus the differences shown in Figure 2 may underestimate the perceptual differences between long-vowel and connected-vowel sequences in terms of auditory streaming.

These data suggest that at least one linguistic and one nonlinguistic factor serve to constrain the identification of rapid vowel sequences. Auditory streaming, a nonlinguistic constraint, affects some sequences more than others. Streaming, however, can be suppressed by interpolating formant transitions between vowel nuclei. When these interconnected sequences are made briefer their identification appears to be less constrained by auditory streaming than by verbal transformations, a resolutely linguistic phenomenon (Warren & Warren, 1970).

SUMMARY AND CONCLUSION

From the results of Experiments 1 and 2 we conclude that certain sequences of four concatenated vowels are more difficult to identify than others, and that the difficulty in identifying these sequences is intimately related to the phenomenon of auditory stream segregation or "streaming." The outcome of Experiment 3 suggests that streaming cannot be effectively suppressed by replacing most of the vowel with silence, but that it can be inhibited by replacing most of the vowel with formant transitions appropriate for stop consonants or with transitions that connect vowel nuclei in an articulatorily reasonable manner. The results of Experiment 4 indicate that the vowel component primarily responsible for streaming is the first formant, but that the second formant may be more important for certain sequences. Finally, from Experiment 5 we conclude that the identification of brief formant-connected sequences is further constrained by the adequacy of the acoustic cues for the perception of vowels as opposed to other segments. In summary, the more repeating vowel sequences resemble connected discourse, the easier they are to identify, and the more linguistic constraints rather than auditory constraints limit the identification of temporal order.

Teleologically speaking, the data suggest a third function of formant transitions in speech. The first function of transitions is to carry phonetic information, and the second is to carry it in such a manner that there is parallel transmission of the phonetic segments (see Liberman et al., 1967). A third function, suggested both by the results of the present studies and by those of Cole and Scott (1973), is to bind together phonetic segments so that at rapid transmission rates the temporal order of speech may be preserved.

REFERENCES

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.

Bregman, A. S., & Dannenbring, G. L. The effect of continuity on auditory stream segregation. *Perception & Psychophysics*, 1973, *13*, 308–312.

Cole, R. A., & Scott, B. Perception of temporal order in speech: The role of vowel transitions. *Canadian Journal of Psychology*, 1973, *27*, 441–449.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. Perception of the speech code. *Psychological Review*, 1967, *74*, 431–461.

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

Neisser, U., & Hirst, W. Effects of practice on the identification of auditory sequences. *Perception & Psychophysics*, 1974, *15*, 391–398.

Orr, D. B., Friedman, H. L., & Williams, J. C. Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology*, 1965, *56*, 148–156.

Thomas, I. B., Hill, P. B., Carrol, F. S., & Garcia, D. Temporal order in the perception of vowels. *Journal of the Acoustical Society of America*, 1970, *48*, 1010–1013.

Warren, R. M., & Warren, R. P. Auditory illusions and confusions. *Scientific American*, 1970, *233*, 30–36.

(Received August 12, 1974)