Identification of Vowel Order: Concatenated Versus Formant-Connected Sequences

M. F. Dorman,‡ James E. Cutting,+++ and Lawrence J. Raphael‡
Haskins Laboratories, New Haven, Conn.

In order for listeners to identify accurately the temporal order of a sequence of four concatenated vowels, each vowel must be between 125 and 200 msec in duration (Warren, 1968; Thomas, Hill, Carrol, and Garcia, 1970). These data fit nicely into a theory of speech perception (Massaro, 1972) that suggests that the perceptual processing of vowels lasts between 120 and 250 msec.

The perception of concatenated speech, however, bears only a slight resemblance to the perception of speech in connected discourse. For example, in discourse listeners can follow speech at rates up to 400 words per min, or approximately 30 phonemes per sec (Orr, Friedman, and Williams, 1965). In marked contrast to the results of Warren (1968) and Thomas et al. (1970), these data suggest that speech signals can be correctly ordered when the average phoneme duration is only 30 to 40 msec. The discrepancy between these two estimates suggested to us that concatenated sequences produce a spuriously high estimate of vowel duration necessary for the temporal ordering of speech. This observation prompted a series of experiments comparing temporal order judgments for concatenated and coarticulated sequences of vowels.

EXPERIMENT I

Method

Warren-type repeating sequences were generated on the Haskins Laboratories' parallel resonance synthesizer. The sequences consisted of long steady-state vowels (V₁) [i, æ, ɔ, u] and of consonant-vowel-consonant (CVC) syllables [bɪ, bæb, bɒb, bub]. All stimuli were 120 msec in duration, well below the critical duration (168 msec) noted by Thomas et al. (1970) for 75 percent correct performance on synthetic vowels. All stimuli had the same fundamental frequency (110 Hz) and overall amplitude contour. Initial and final formant transitions in the CVCs were 45 msec in duration, leaving 30 msec of steady-state vowel. Stimuli within the same class (V₁ and CVC) were permuted in the six possible orders. These sequences were recorded on audio tape with 10 msec between successive items. 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), remained at that

‡ Also Herbert H. Lehman College of the City University of New York.
+++ Also Yale University, New Haven, Conn.

maximum for 10 sec, and then decreased to its original low volume during the final 5 sec period. Stimulus class and sequence orders were randomized for presentation to the 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 sequences and CVC sequences, and were instructed to report the identity of the vowels in the order that they were heard (disregarding the /b/s in the CVC stimuli).

Results and Discussion

Table 1 shows the averaged performance of the listeners for \( V_1 \) and CVC stimuli for each of the six orders. Notice that in terms of performance summed over all sequence orders, there was no significant difference between the two classes of stimuli: \( V_1 \) and CVC stimulus orders were correctly identified on 67 and 70 percent of all trials, respectively. This result is deceptive, however, since it sums over sequence orders with quite varied performance levels. Two orders, number 2 [i, æ, u, ɔ] and number 4 [i, ɔ, u, æ] were more difficult to identify than the other four. The results of these two orders yielded the most interesting pattern: \( V_1 \) sequences were identified on only 51 percent of such trials, and CVC sequences were identified on 64 percent. This pattern occurred against a background of small differences between the two stimulus classes on the other four sequence orders: 75 percent correct for \( V_1 \) stimuli, and 73 percent for CVCs. Listeners readily volunteered that sequence orders number 2 [i, æ, u, ɔ] and number 4 [i, ɔ, u, æ] were more difficult to identify than the others, describing the difficulty in terms of the sequences "flying apart." We recognized this as the hallmark of "auditory streaming."

Bregman and Campbell (1971) have reported that when listeners are presented a repeating sequence of six brief (100 msec) tones which alternate between high and low frequencies, listeners are unable to report correctly the high-low sequence, reporting instead two streams of tones, one containing the high tones and the other containing the low tones. Within a stream, ordering is reasonably
accurate (73 to 79 percent), however between streams ordering is no better than chance. Bregman and Campbell have termed this phenomenon "primary auditory stream segregation."

The perceptual experience in the present study of the vowel sequences "flying apart" on sequences 2 and 4 appears identical to the auditory stream segregation of Bregman and Campbell's nonspeech auditory signals. Although sequences 2 and 4 were characterized by physically separated [i]-[u] and [æ]-[ɔ] pairs, listeners reported hearing each as a unit pair. This phenomenon may be accounted for in terms of perceptual streaming of the first formants, the most prominent and lowest-frequency component of the four vowels. The first formant (F1) frequency value for both [i] and [u] was 286 Hz, whereas for [æ] and [ɔ] it was 666 Hz and 614 Hz, respectively. Since [i] and [u] have the lowest-frequency first formants, these vowels appear to be heard as one stream, and [æ] and [ɔ] with higher first formants appear to form a separate stream, as suggested in the top panel of Figure 1. Only sequence orders 2 and 4 meet the requirement of having alternating high and low first formants, and indeed these are the orders that were most difficult for listeners to identify.

A second important observation is that the sequence order of 30 msec vowels in the context of initial and final [b] could be identified at least as accurately as the 120 msec vowels. Since 30 msec is far below the vowel duration necessary for accurate temporal ordering of concatenated vowels (Thomas et al., 1970), the formant transitions in the CVC sequences may act in a manner similar to silence in facilitating recognition of vowel sequence order. Warren (1968), for example, has reported that, although sequence orders of four concatenated 200-msec vowels are very difficult to perceive, the same orders with 150 msec vowels separated by 50 msec of silence are relatively easy to perceive.

Since the results of the present study suggest that certain sequence orders are more difficult to identify than others, Experiment II was designed, in part, to observe such differences in greater detail. In addition, Experiment II was designed to compare the relative contribution of transitions and of silence to the perception of temporal order.

EXPERIMENT II

Method

The V1 and CVC stimuli from the previous experiments were used again. In addition, short, steady-state vowel stimuli (V8) were synthesized. They were identical to the V1 stimuli in all respects except duration. Whereas the long vowels were 120 msec in duration, the short vowels were only 30 msec long, and thus identical to the steady-state vowel portion of the CVC stimuli. The other 90 msec of the stimulus was replaced by silence. Again, all stimuli within a class were permuted in the six possible sequence orders, but the two most difficult orders (numbers 2 and 4) were represented twice as often as the other four. Schematic spectrograms of the V1, V8, and CVC stimuli in the order [i, ɔ, u, æ] are shown in Figure 1. Each sequence was recorded in the same fashion as in Experiment I, and class of stimuli and sequence order were randomly intermixed.

Eight students at Herbert Lehman College of the City University of New York and three staff members from Haskins Laboratories served as listeners. Stimuli were reproduced for the Lehman College listeners on a Revox 1122 tape recorder via an AR-4x loudspeaker, and for the Haskins Laboratories listeners on the same apparatus as in Experiment I.
Figure 1: Schematic spectrograms of long-vowel ($V_l$), short-vowel ($V_s$), and consonant-vowel-consonant (CVC) sequences.
Results and Discussion

As shown in Table 2, V₁ and Vₛ sequences were considerably more difficult to order than the CVC sequences, with their respective average performance levels at 52, 47, and 71 percent. CVC sequence orders were significantly easier to

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>V₁</th>
<th>Vₛ</th>
<th>CVC</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) i æ c u</td>
<td>88</td>
<td>50</td>
<td>88</td>
<td>75</td>
</tr>
<tr>
<td>*2) i æ u c</td>
<td>44</td>
<td>38</td>
<td>78</td>
<td>53</td>
</tr>
<tr>
<td>3) i c æ u</td>
<td>50</td>
<td>13</td>
<td>50</td>
<td>38</td>
</tr>
<tr>
<td>*4) i u æ æ</td>
<td>39</td>
<td>50</td>
<td>67</td>
<td>52</td>
</tr>
<tr>
<td>5) i u æ c</td>
<td>63</td>
<td>75</td>
<td>88</td>
<td>75</td>
</tr>
<tr>
<td>6) i u c æ</td>
<td>50</td>
<td>63</td>
<td>50</td>
<td>54</td>
</tr>
</tbody>
</table>

*represented twice as often as other orders.

identify than those of either the V₁ stimuli (T(8) = 3, p < .05) or the Vₛ stimuli (T(8) = 5, p < .05). The differences are even more striking for the two difficult sequence orders, numbers 2 and 4: such V₁ sequences were correctly identified on only 42 percent of all presentations, Vₛ sequences were identified on 44 percent, and CVC stimuli on 73 percent. For the other sequences (1, 3, 5, and 6) temporal-order accuracy was 63, 50, and 69 percent for the V₁, Vₛ, and CVC sequences, respectively. There were no significant differences among the sequences of V₁ and Vₛ stimuli. There were also no systematic differences between the Haskins and Lehman subjects.

The perceptual advantage of the CVC sequences over the two classes of vowels stems primarily from performances on the sequence numbers 2 and 4. This outcome suggests that the reason there was no significant difference between the CVC and V₁ stimuli in Experiment I was a ceiling effect induced by the over-representation of the easier sequence numbers 1, 3, 5, and 6. The perceptual advantage of the CVC sequences over the Vₛ sequences suggests that tracking formants between vowels is more effective than silence in reducing auditory streaming. Such an outcome is encouraging since few silent intervals occur in running speech, yet correct phoneme order is effortlessly extracted.

The superiority of the CVCs over both types of vowel sequences suggests further that there may be other ways of perceptually "gluing" the vowels together. Experiment III was designed, in part, to determine if streaming could be overcome through the use of long, continuous transitions between vowel nuclei. This tactic has proved useful in limiting primary auditory stream segregation in sequences of pure tones (Bregman and Dannembring, 1973). Experiment III was also designed to determine if streaming is suppressed by all transitions, or only by transitions that are phonetically possible.

221
EXPERIMENT III

Method

The \( V_1 \) and CVC stimuli were used again. In addition, two other sets of stimuli were generated. Both sets contained 30-msec steady-state vowel segments corresponding to the vowels \([i, \, a, \, o, \, u]\) and both had initial and final formant transitions. In one set the transitions were context-dependent, gliding gradually over the course of 90 msec from the steady-state formant values of one vowel into the succeeding vowel. Because of the connecting transitions, they are termed \( V_t \) stimuli. The second set, termed \( C'VC \) stimuli, contained most of the features of CVCs except that the formant transitions were turned upside down; that is, instead of all transitions gliding upwards 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 stimulus orders were selected: orders 2 and 4 to optimize error probability, and order number 1 for comparison purposes. \( V_1, \ V_t, \ CVC, \) and \( C'VC \) sequences of \([i, \, o, \, u, \, a]\) are shown in Figure 2. Class of stimuli and sequence order were randomized and recorded on audio tape.

The listeners were nine Lehman College undergraduate students. Stimuli were reproduced on a Revox 1122 tape recorder via an AR-4x loudspeaker. In all other respects the procedure was the same as in the two previous studies.

Results and Discussion

As shown in Table 3, there was a large difference between the accuracy of sequence identification for the two types of vowel stimuli. \( V_1 \) sequence orders were identified on only 30 percent of all presentations, whereas \( V_t \) orders were identified on 65 percent. All subjects demonstrated this difference (\( T(9) = 0, \ p < .01 \)). The CVC sequences were again identified more accurately than the long vowels (\( T(9) = 0, \ p < .01 \)), and again all listeners demonstrated this effect. There was no difference between CVC and \( V_t \) sequence-order identification. \( C'VC \) sequence orders were essentially incomprehensible, and were identified at a chance performance level.

<table>
<thead>
<tr>
<th></th>
<th>( V_1 )</th>
<th>( V_t )</th>
<th>CVC</th>
<th>( C'VC )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) i a e c u</td>
<td>50</td>
<td>55</td>
<td>44</td>
<td>11</td>
</tr>
<tr>
<td>2) i a e u c</td>
<td>6</td>
<td>78</td>
<td>83</td>
<td>17</td>
</tr>
<tr>
<td>4) i o u a e</td>
<td>33</td>
<td>61</td>
<td>55</td>
<td>11</td>
</tr>
<tr>
<td>Average</td>
<td>30</td>
<td>65</td>
<td>61</td>
<td>13</td>
</tr>
</tbody>
</table>

CONCLUSIONS

Two trends are prominent in the data of the present study. First, the employment of gradual transitions between the 30-msec vowel nuclei was successful
Figure 2: Schematic spectrograms of long-vowel ($V_l$) and vowel-with-transition ($V_t$) sequences; consonant-vowel-consonant (CVC) sequences; and sequences with phonetically impossible transitions (C'VC').
in "gluing" back together the previously "streamed" vowel sequence orders. Such transitions are reasonable in an articulatory sense, just as the transitions in the CVC sequences are also reasonable. This points out the second trend: only those transitions that are phonetically possible were successful in limiting auditory streaming. Articulatorily impossible transitions such as those in the C'Ve' stimuli failed to limit streaming. In fact, they decreased the probability of perceiving correct stimulus order.

From the results of Experiments I-III, we conclude (a) that certain sequence orders of four vowels are more difficult to perceive than others, (b) that the difficulty in perceiving these sequence orders is intimately related to the phenomenon of auditory stream segregation, (c) that streaming cannot be eliminated by replacing most of the vowel with silence, but that it can be virtually eliminated by replacing most of the vowel with formant transitions appropriate for the stop consonant [b] or with formant transitions that link the vowel nuclei with one another, and (d) that the suppression of auditory streaming is possible only through the use of transitions corresponding to gestures that could be articulated. In other words, the more the repeating sequences resemble connected discourse, the more facile listeners are at identifying temporal order. Phonemes in connected discourse do not stream precisely because they are coarticulated and not concatenated.

REFERENCES

Bregman, A. S. and J. Campbell. (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J. Exp. Psychol. 89, 244-249.


