On the contribution of speech segments to the identification of Swedish consonant phonemes

Öhman, S.

journal: STL-QPSR
volume: 2
number: 2
year: 1961
pages: 012-015

http://www.speech.kth.se/qpsr
II. SPEECH PERCEPTION

A. ON THE CONTRIBUTION OF SPEECH SEGMENTS TO THE IDENTIFICATION OF SWEDISH CONSONANT PHONEMES

There are three principal perceptual problems connected with acoustic features and segments. Given a sound segment:

1. In what ways and to what extent does it contribute to the identification of the phonemes in front of it, behind it, and of the phoneme of which it is, acoustically, a part?
2. What role do the abrupt acoustic segment boundary phenomena play in perception?
3. Which of the distinctive sound qualities characterizing the phonetic system of the language are, as it were, in action in the segment under scrutiny, and to what extent?

In order to gain some insight into these questions an investigation is being undertaken, some results of the preliminary part of which are reported on here. Briefly speaking, the experiment consists in presenting listeners with a set of VCV nonsense words which commute the C and where variously large parts are cut away by means of an electronic gate and then in studying the differences in responses relative to the differences in place for cut.

The following disyllabic nonsense words, comprising long consonants between short vowels, were recorded anechoically on tape:

/appa  atta  ar\textsuperscript{\textregistered}a  akka
abba  adda  ar\textsuperscript{\textregistered}a  agga
amma  anna  ar\textsuperscript{\textregistered}a  anga
affa  assa  a\textsuperscript{\textregistered}ja  atja
svva  ar\textsuperscript{2}r\textsuperscript{2}a  a\textsuperscript{\textregistered}ja
alla  arra  ahha/

This series includes all Swedish consonants of which the following are referred to as supradentals: \( r\textsuperscript{\textregistered} = [\textregistered] \), \( r\textsuperscript{\textregistered}d = [\textregistered \textregistered] \), \( r\textsuperscript{\textregistered}m = [\textregistered] \), \( a\textsuperscript{\textregistered} = [\textregistered] \), \( r\textsuperscript{\textregistered} = [\textregistered] \). The sign \( r\textsuperscript{\textregistered} \) stands for a Stockholmian allophone of the /r/-phoneme which is also represented by the more...
standard tremulant apical r₁ = [r]. The sign t₂ = [q]. These words are pronounced with the normal grave (disyllabic) accent by one male speaker. Average F-pattern values for this speaker’s /a/:s are - F₁ = 750 c/s, F₂ = 1200 c/s, F₃ = 2600 c/s, and F₄ = 3250 c/s.

Spectrograms from each of the words were prepared, the left boundaries and the right boundaries were determined (i.e., the time points where the initial vowel ended and the intervocalic consonant started, and where the consonant ended and the final vowel started, respectively. This was done in accordance with current praxis, and the final half of the initial vowel, all of the consonant, and the initial half of the final vowel were divided into immediately adjacent 20 msec long segments by marking off points along the time scales of the spectrograms.

With this segmentation in mind the following question could be put:

Presenting to unbiased listeners first one part of a word (say, the initial part up to segment No. 10) and then the same part of the same word but this time eked out with one 20 msec long segment (i.e., up to segment No. 11), what are the differences in perceptual response? Obviously one can make this experiment both with and against the direction of time, i.e., one can complement the above mentioned test with a test where one starts with the final half of a word, then eked this out with a 20 msec long segment at the front end, then eventually determines the differences in response. In this way one can obtain, among other things, two curves for each VCV word, one showing percent correct identification as a function of time when the stimuli are eked out in the forward direction of time, and one showing percent correct identification of the consonant as a function of time when the stimuli are eked out in the backward direction. This report comments only on some results from the "forward" test.

The segmentation of the words was made by means of an electronic gate. Tape loops were prepared, one for each of the VCV words, which when played over and over again allowed continuous inspection of the intensity-time function of the speech signal on a cathode ray oscilloscope. This showed the complete signal on one sweep and
the gated and synchronized version of the same signal on the other sweep. When the gate was adjusted to the predetermined time positions the split signal was recorded on a double track tape recorder, the first half of the word on one track and the other half of the word on the other track. Then the gating point was moved 20 msec to the right, a new recording was made and so on with all the words.

The stimuli were now gathered from one track of the last recording. They all had the character of VC syllables now, since all word parts containing any fraction of the second vowel were excluded. They were arranged in random order and presented to 18 listeners (both male and female) via ear-phones. The listeners were instructed to identify the final consonants by means of writing down one of 21 letters on a response sheet. They had a short period of training for acquisition of a correct usage of the response symbols.

Results

We cannot give a detailed account of the results of the test here, which are selectively presented in the form of confusion matrices and graphs at the end of this paragraph. Instead, we simply describe the general contents of these figures.

The test can be regarded as a number of tests run simultaneously. One such part is the test which consists of stimuli cut two steps to the left of the segment containing the physically predetermined (left) boundary, i.e., about 40 msec before the physical boundary. Of course, each VCV word provides one such stimulus. Hence we can set up a 22 x 22 matrix showing consonants presented in the rows and letters responded by eighteen listeners to each of these consonants in the columns. The time parameter is fixed for this matrix to two steps to the left of left physical boundary. The matrix is shown as Fig. II-1. In the same way matrices are prepared for stimulus cuts corresponding to "on the left physical boundary" (Fig. II-2), "two steps to the right of left physical boundary" (Fig. II-3), and "two steps to the left of right physical boundary" (Fig. II-4).
<table>
<thead>
<tr>
<th>t</th>
<th>r</th>
<th>t</th>
<th>k</th>
<th>b</th>
<th>d</th>
<th>r</th>
<th>g</th>
<th>m</th>
<th>n</th>
<th>n</th>
<th>g</th>
<th>f</th>
<th>s</th>
<th>sj</th>
<th>tj</th>
<th>v</th>
<th>r2</th>
<th>j</th>
<th>l</th>
<th>r1</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sj</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tj</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>9</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>r1</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Rows = stimuli presented  Columns = responses from 18 listeners

Fig. II-1 Confusion matrix for stimuli cut two steps to the left of left physical boundary.
Fig. 11-2 Confusion matrix for stimuli cut on the left physical boundary.
Rows = stimuli presented  
Columns = responses from 18 listeners

Fig. II-3 Confusion matrix for stimuli cut two steps to the right of left physical boundary.
Rows = stimuli presented  
Columns = responses from 18 listeners

Fig. II-4 Confusion matrix for stimuli cut two steps to the left of right physical boundary.
Fig. II-5 shows the confusions among consonant classes for each of the four above mentioned cut points. Here, e.g., "voice/unvoiced" means that out of the total number of responses to voiced or unvoiced sounds, so and so many percent were correct in using either a "voiced letter" when a voiced sound was presented or an "unvoiced letter" when an unvoiced sound was presented. "Occlusive" means that out of the total number of responses to occlusive sounds, so and so many percent responded with "occlusive letters", so and so many with "nasal letters" and so on.

The following table shows the principles of classification:

**Voiced:** /b d r r g m n r n ng v r₂ j l ř₁/
**Unvoiced:** /p t ɾ t k f s sj tj h/  
**Occlusive:** /p t ɾ t k b d r d g/  
**Nasal:** /m n r n ng/  
**Liquid or h:** /l r₁ h/  
**Fricative:** /f s sj tj v r₂ j/  
**Labial:** /p b m f v/  
**Dental:** /t d n s l/  
**Supradental:** /ɾ t ɾ ɾ r s j r₂ r₁/  
**Palatal or h:** /k g ng j ɾ j h/  

The results indicate that for most words (not for appa, atta, and arta) it is quite possible to spot a segment where "(almost) full identification" occurs for the first time, i.e., a segment which perhaps could be labeled a perceptual boundary point. Presumably there is on the average not only one such point for each word. For the identification of an /m/ in /amma/, for instance, there are two critical points; one about at the middle of the /m/ and one at the beginning of the final /a/. However, comparing the first critical points with the points corresponding to the physically determined (left) boundaries, it is possible to extract two differences between the two points. One, Δt, for time, and another, Δi, for percent identification. These differences are plotted in Fig. II-6. As can be seen, the physical boundaries are not coincident with the "perceptual boundaries" as defined above. However, they appear to be systematically interrelated.

S. Ohman
Fig. II-5 Confusion among classes of consonants. The matrices show the percentages with which consonants in the rows were confused with consonants in the columns.
\( \Delta t = \text{Distance in msec between physical boundary as seen on spectrogram and perceptual boundary. The latter is taken to be where maximum average identification of the sound occurs for the first time.} \)

**Fig. II-6** Examples illustrating how the graph may be read:

1. The consonant /m/ (18, 41) was fully identified only 18 msec. later than left physical boundary, and at the latter point (i.e., at the physical boundary) identification was 41% worse than maximum for that sound.

2. Consonant /j/ (-30, -2) had full identification 30 msec. before physical boundary. At the latter point identification was slightly higher than average max. (within one standard deviation unit).

\( r_1 \) is ordinary apically tremulant (r), and 
\( r_2 \) is the voiced fricative alveolar Stockholmian r-allophone.

For consonants /p t rt/ neither average max. values nor first point of max. i could be determined.