On the Identification of Place and Voicing Features in Synthetic Stop Consonants

David B. Pisoni\(^+\) and James R. Sawusch\(^++\)

Two models of the interaction of phonetic features in speech perception were used to predict subjects' identification functions for a bidimensional series of synthetic consonant-vowel syllables. The stimuli varied systematically in terms of the acoustic cues underlying the phonetic features of place of articulation and voicing. Model I assumed that phonetic features are additive and are processed independently in perception. Model II assumed that the phonetic features interact and are not processed independently. The fit of Model II to the bidimensional series data was better than the fit of Model I, suggesting that the phonetic features of place and voicing in stop consonants are not processed independently but rather show a mutual dependency.

Theoretical accounts of speech sound perception have frequently proposed some type of articulatory-motor involvement during perceptual processing (Liberman, 1957; Stevens, 1960; Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens and Halle, 1967). One reason for this may be that research on speech sound perception has drawn its descriptive categories from the account of speech production offered by phoneticians. Thus, the articulatory dimensions that distinguished different classes of speech sounds in production served as the basis for uncovering the acoustic cues that distinguish different speech sounds in perception. Spectrographic analysis and perceptual experiments revealed that the sounds of speech were not arrayed along a single complex dimension but could be specified in terms of a few simple and independent dimensions (Gerstman, 1957; Liberman, 1957). Acoustic dimensions were found in early experiments with synthetic speech to provide distinctions in perception corresponding to the articulatory dimensions of speech production, suggesting that perceptual and articulatory dimensions of speech may be intimately linked (Delattre, 1951; Liberman, 1957).

\(^+\)Indiana University, Bloomington, and Haskins Laboratories, New Haven, Conn.

\(^++\)Indiana University, Bloomington.

Acknowledgment: This research was supported in part by a PHS Biomedical Sciences Grant (S05 RR 7031) and PHS Training Grant (T01 MH 11219-04) to Indiana University and in part by a grant from the National Institute of Child Health and Human Development to Haskins Laboratories. We wish to thank M. Studdert-Kennedy and R. M. Shiffrin for comments and suggestions on the manuscript.


65
Two articulatory features to receive a great deal of attention in the description of stop consonant production are place of articulation and voicing. Both these features have fairly well defined acoustic properties which presumably mirror the differences in production (Delattre, 1951). For example, the feature of place of production refers to the point of constriction in the vocal tract where closure occurs. The acoustic cues that underlie the place feature in consonant-vowel (CV) syllables are reflected in the formant transitions into the following vowel, particularly the direction and extent of the second and third formant transitions (Liberman et al., 1967). In contrast, the voicing feature is related to the presence or absence of periodic vibration of the vocal chords. The acoustic cues that underlie the voicing feature in stop consonants in initial position are reflected in terms of the relative onset of the first formant transition (i.e., F1 "cutback") and the presence of aspiration in the higher formants (Liberman, Delattre, and Cooper, 1958). This compound acoustic cue has been called "voice-onset time" (VOT) by Lisker and Abramson (1964) and corresponds to the time interval between the release from stop closure and the onset of laryngeal pulsing.

Figure 1 presents schematized spectrographic patterns which show the acoustic cues for place and voicing features for the CV syllables /ba/, /da/, /pa/, and /ta/. There is a relatively simple relation between articulatory features of place and voicing and their respective acoustic cues when the vowel is held constant (see also Liberman, 1970). Consonants within a particular row share voicing: /ba/ and /da/ are voiced, /pa/ and /ta/ are voiceless. The major acoustic cue for voicing in these syllables is the cutback or elimination of the initial portion of the first formant. Consonants within a particular column share place of production: /ba/ and /pa/ are bilabial stops, /da/ and /ta/ are alveolar stops. The primary acoustic cue for place is the direction and extent of the second and third formant transitions.

Several perceptual experiments employing stop consonant-vowel syllables have concluded that the features of place and voicing are processed independently of each other. For example, Miller and Nicely (1955) analyzed the perceptual confusions among 16 CV syllables presented to listeners under various signal-to-noise ratios and filtering conditions. They computed the sum of the information transmitted by the features separately and in combination. Since the two values were approximately equal, they concluded that the features used in their analysis were mutually independent. Among these features were place and voicing. As a part of a larger investigation of dichotic listening, Studdert-Kennedy and Shankweiler (1970) reached the same conclusion by a similar analysis of place and voicing confusions among stop consonants.

These studies imply that features are extracted separately during early perceptual processing and are later recombined in response. Figure 2 represents a simplified block diagram of this process. The output of the auditory analysis is a set of acoustic cues \{c_i\}. These cues are combined and from them a set of phonetic features \{f_j\} is recognized. Finally, the phonetic features are combined to yield the perception of the phonetic segment. Together, stages 2 and 3 form what Studdert-Kennedy (1973) has described as the "phonetic" stage of processing. We assume that phonetic features are recognized or identified in short-term memory (STM) when the auditory patterns derived from the acoustic cues have made contact with some representation generated from synthesis rules residing in long-term memory. We assume that abstract phonetic features have an articulatory rather than acoustic reality in STM although we will not try to justify this assumption at present.
Figure 1: A schematized sound spectrogram of the syllables /ba/, /da/, /pa/, and /ta/ as used in the present experiment.
Figure 2: A block diagram of stages of perceptual analysis for phonetic segments. The input is the acoustic waveform and the output is a sequence of phonetic segments.
To say that phonetic features are independent implies independence of processing at all three stages of Figure 2. That is, the acoustic cues are extracted from the acoustic waveform separately and independently of each other in stage 1. Then the phonetic features are extracted separately and independently from the acoustic cues in stage 2. Finally, the phonetic features are combined separately and independently of each other in stage 3, resulting in a particular phonetic segment.

The independence of phonetic feature processing in stage 3 may be described quantitatively by a simple linear or additive model. The phonetic features of place \( f_1 \) and voicing \( f_2 \) that are output from stage 2 are weighted separately and then added together in stage 3. Equation 1 expresses this concept algebraically:

\[
X = a_1 f_1 + a_2 f_2
\]  
(1)

Here, \( f_1 \) is the amount of the place feature of stimulus \( X \) output from stage 2, and \( a_1 \) is its associated weight. Similarly, \( f_2 \) is the amount of the voicing feature of \( X \) output from stage 2, and \( a_2 \) is its associated weight. Since these two features are sufficient to distinguish among the four stops \( b, d, p, \) and \( t \), we will ignore other phonetic features as being redundant and nondistinctive.

However, evidence for nonindependence of phonetic features, in particular the features of place and voicing, has also been presented by several investigators. This nonindependence could come at any of the three levels mentioned. For example, Haggard (1970) put the dependence relationship of the features of place and voicing in the second stage, where phonetic features are extracted. In the model of Haggard (1970), the listener's decision on the voicing feature is partly determined by his prior decision on the place feature.

Lisker and Abramson (1964, 1967) reached the corresponding conclusion upon examination of their production data for stop consonants in initial position. The voicing feature as reflected in VOT depends on the feature of place of production; the VOT lag at the boundary between voiced and voiceless stops increases as place of production moves further back in the vocal tract (i.e., from \(/ba/\) to \(/da/\) to \(/ga/\)). Given the anatomical and physiological constraints on speech production, this position is a priori more plausible.

This particular concept of nonindependence, where the feature of voicing partly depends on the feature of place, may also be expressed algebraically. We will again assume independence of processing in stages 1 and 2. The nonindependence in stage 3 may be expressed as:

\[
X = a_1 f_1 + a_2 f_2 + b(1-f_1)f_2
\]  
(2)

Here, \( a_1, f_1, a_2, \) and \( f_2 \) are the same as in equation 1. However, the constant \( b \) represents the weight given to the interaction term of place and voicing \([(1-f_1)f_2]\).

The purpose of the present experiment was to reexamine the identification of place and voicing features in stop consonants and to determine by means of a new experimental paradigm whether these two phonetic features (i.e., place and voicing) are combined additively or nonadditively in stage 3 as shown in Figure 2.
Stimuli for the experiment were three sets of synthetic speech sounds that varied systematically in the acoustic cues underlying the two phonetic features of place and voicing. One series of stimuli varied in the acoustic cues that underlie the phonetic features of place, while holding the voicing feature constant (/ba/ to /da/ with VOT at 0 msec). A second series varied the acoustic cues underlying the phonetic feature of voicing, while holding the place feature constant (/ba/ to /pa/ with F2 and F3 always rising). The final series varied the acoustic cues underlying both phonetic features simultaneously (/ba/ to /ta/). These three sets of speech sounds were presented separately to listeners for identification into the categories /ba/-/da/, /ba/-/pa/, and /ba/-/ta/, respectively. The use of synthetically produced stimuli made it possible to control experimentally the correlation between place of production and voicing that Lisker and Abramson (1964, 1967) had found in natural speech.

Our principle aim was to determine whether the probabilities of identification along the bidimensional continuum (/ba/ to /ta/) could be predicted from some combination of the probabilities along the separate unidimensional series. We consider below two possible models of ways these separate features might be combined in the bidimensional case. Both Model I and Model II are concerned with the manner in which phonetic features are combined in phonetic perception. All processing up to stage 3 of Figure 2 is assumed to be independent according to the definition of independence for these stages given previously. We also assume that processing in stages 1 and 2 takes place in parallel and is automatic in the sense that Ss do not have control over these stages of perceptual processing. (See also Shiffrin and Geisler, 1973, and Shiffrin, Pisoni, and Castenada-Mendez, in press.)

Model I: Linear Combination of Phonetic Features

Hereafter, if a S identified a stimulus as /ba/ it will be denoted B and likewise, /da/ as D, /pa/ as P, and /ta/ as T. In the /ba/ to /da/ series only the acoustic cues underlying the phonetic feature of place of articulation were varied. Since processing in stage 2 is assumed to be independent (i.e., separate for different phonetic features), the only variation in the output of stage 2 on the /ba/ to /da/ (place) series should be in feature $f_1$, the phonetic feature of place of articulation. Accordingly, since the only variation in the input to stage 3 is in $f_1$, the output of stage 3 (a phonetic segment) is assumed to vary directly with the input ($f_1$) and thus accurately reflect $f_1$. However, due to noise in the acoustic waveform and in the first two stages of processing, the outputs of stage 2 are assumed to be probabilistic in nature. Thus, Ss' judgments of the stimuli from the /ba/ to /da/ series (the probability of responding D to a stimulus, $Pr[D]$) may then be construed as accurately reflecting the input ($F_1$) to stage 3. Similarly, $Pr[P]$ from the /ba/ to /pa/ (voicing) series may be construed as accurately reflecting the input of the voicing feature ($f_2$) to stage 3. Now, we can represent $f_1$ and $f_2$ from equations 1 and 2 as follows:

$$f_1 = Pr[D] \text{ on the /ba/-/da/ series (PLACE)} \quad (3)$$

$$f_2 = Pr[P] \text{ on the /ba/-/pa/ series (VOICING)} \quad (4)$$

Substituting equations 3 and 4 into equation 1 we obtain equation 5:

$$Pr[T] = a_1 Pr[D] + a_2 Pr[P] \quad (5)$$
Pr[T] in equation 5 represents the probability of a T response on the bidimensional /ba/ to /ta/ series.

One additional assumption will be made. This is shown in equation 6:

\[ a_2 = 1 - a_1 \text{ where } 0 \leq a_1 \leq 1 \]  

(6)

This constraint is placed on \( a_1 \) and \( a_2 \) so that \( Pr[T] \) will equal one when both \( Pr[D] \) and \( Pr[P] \) are equal to one. Since only one parameter is being used, we delete the subscript from the constant \( a \).

If we now combine equations 5 and 6 and delete the subscript on the constant \( a \) we obtain equation 7:

\[ Pr[T] = a Pr[D] + (1-a) Pr[P] \]  

(7)

Equation 7 represents Model I. This model assumes independence of the features of voicing and place. If we estimate parameter \( a \) from the data by the method of least squares, then Model I can be used to predict the bidimensional /ba/ to /ta/ identification function based on the unidimensional /ba/ to /da/ and /ba/ to /pa/ data.

Model II: Nonlinear Combination of Phonetic Features

A development similar to that given for Model I may be applied to equation 2. If we combine equations 2, 3, 4, and 6, we obtain equation 8:

\[ Pr[T] = a' Pr[D] + (1-a') Pr[P] - b(1-Pr[D]) Pr[P] \]  

(8)

Here, \( a' \) is used to distinguish this parameter from the parameter \( a \) of Model I. A major disadvantage of equation 8 is that it requires two different parameters, \( a' \) and \( b \) to be estimated from the data.

Equation 8 assumes that Ss employ information about both phonetic features, place and voicing, to make their decision on the /ba/ to /ta/ series. However, either of these features alone may be sufficient for a S to distinguish between /ba/ and /ta/ when only two response categories are permitted. For example, a S could identify /ba/ and /ta/ on voicing alone (i.e., if voiced, respond /ba/; if voiceless, respond /ta/) or on place alone (i.e., if bilabial, respond /ba/; if alveolar, respond /ta/). Since these stimuli differ in both voicing and place, Ss may use only one of these features in their decision. However, it is also possible that a particular decision on one feature necessarily entails a particular decision on the other. This is even quite likely considering the constraints on production. In production, a shift in place of articulation entails a shift in VOT, but not vice versa. On the other hand, in perception, the shift in VOT may serve as a cue to a shift in place of articulation.

Previous investigators have found that decisions based on the voicing feature are more consistent and, in some sense, easier than decisions based on other features, including place (Miller and Nicely, 1955; Studdert-Kennedy and Shankweiler, 1970; Shepard, 1972). One reason for this finding may be the multiplicity of cues to the voicing feature (Liberman et al., 1958; Lisker and Abramson, 1964; Summerfield and Haggard, 1972), as compared with the relatively
restricted number of cues to the place feature. If subjects were to use only one feature, it seems likely that they would use the feature of voicing for the /ba/ to /ta/ series. We can operationalize this assumption by setting $a'$ from equation 8 to zero. This means that $1-a'$ will be 1 and that the probability of a /ta/ response (Pr[T]) will be the result of the amount of the voicing feature present minus the interaction of place and voicing. This is summarized in equation 9, which represents Model II:

$$Pr[T] = Pr[P] - b(1-Pr[D]) Pr[P]$$

(9)

The term representing the dependence of voicing on place (1-Pr[D]) Pr[P] has been operationalized this way to insure that Pr[T] does not become negative or greater than one.

Model II can be used to predict the bidimensional /ba/ to /ta/ series from the /ba/ to /da/ and /ba/ to /pa/ data by estimating the parameter $b$ with the method of least squares. By setting parameter $a'$ to zero, Model II assumes that Ss categorize the stimulus as either /ba/ or /ta/ on the basis of the voicing feature alone. Thus, parameter $b$ may be used as an estimate of how much a S's decision on the voicing feature depends upon the place information in the stimulus.

**METHOD**

**Subjects.** Subjects were nine students in introductory psychology, participating as a part of the course requirement. Each S was a native American speaker of English, right-handed, and reported no history of a speech or hearing disorder.

**Stimuli.** The three synthetic speech syllable series were /ba/ to /da/, /ba/ to /pa/, and /ba/ to /ta/. Each series contained 11 stimuli. The /ba/ to /da/ series varied in the initial frequencies of the second and third formant transitions. The second formant varied from an initial value of 1,859 Hz (/ba/) to an initial value of 3,539 Hz (/da/) in ten equal steps. The /ba/ to /pa/ series varied in VOT from 0 msec VOT (/ba/) to a +50 msec VOT (/pa/) in 5 msec steps. Aspiration replaced the harmonics in the second and third formant transitions for the duration of the F1 cutback. The /ba/ to /ta/ series combined the two components in a one-to-one fashion, resulting in the third 11-stimuli sequence. All stimuli were of 300 msec duration with a 50 msec transitional period followed by a 250 msec steady-state vowel (/a/). The three series of synthetic stimuli were prepared on the speech synthesizer at Haskins Laboratories and recorded on magnetic tape.

**Procedure.** The experimental tapes were reproduced on a high quality tape recorder (Kenwood AG-500) and were presented binaurally through Telephonic (TDH-39) matched and calibrated headphones. The gain of the tape recorder playback was adjusted to give a voltage across the headphones equivalent to 80 db SPL re 0.0002 dyn/cm$^2$ for the steady-state calibration vowel /a/.

On each tape 5S heard 10 presentations of each of the 11 stimuli in random order with 4 sec between stimuli. Ss were run in two groups, 5 Ss in the first group and 4 Ss in the second. Each group heard each tape three times, resulting in 30 judgments of each stimulus for each S. In addition, the /ba/ to /ta/ tape was presented twice more with a different set of instructions. The order of tape presentation was randomized with one group hearing the /ba/ to /da/ tape first.
For each of the tapes Ss were told that they would hear synthetic speech syllables and they were to identify them as /ba/ or /da/, /ba/ or /pa/, /ba/ or /ta/. Ss were told to record their identification judgment of each stimulus by writing down the initial stop consonant in prepared response booklets.

RESULTS AND DISCUSSION

The identification probabilities for the /ba/ to /da/ (place) and /ba/ to /pa/ (voicing) series were in accord with previous experiments. All the stimuli at one end of the series were consistently categorized one way and all the stimuli at the other end were consistently categorized the other way. There were a few transition stimuli (generally one or two in the middle of the series) which were categorized both ways at a near chance (.5) level. Data from two Ss were eliminated from subsequent analyses since they responded to the /ba/ to /da/ series at a chance level throughout. (One S came from each of the groups.) Identification functions from these two series for a typical S (S number 1) are shown in Figures 3A and 3B. The /ba/ to /ta/ identification function for the same S is shown in Figure 3C.

In order to estimate the weighting factor (a) from Model I for each S, a was allowed to vary from 0.0 to 1.0 in increments of .02. The squared error between the predicted and observed identification functions was then calculated for each value of a. The value which resulted in the minimum squared error for each S was chosen as the best estimate of a. These values of parameter a and their associated squared errors are shown in Table 1. In six out of seven Ss the proportion of the variance accounted for by the predicted values exceeded 86 percent. The mean proportion of variance accounted for over all Ss by Model I was 89.7 percent.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Constants a</th>
<th>1-a</th>
<th>Minimum Squared Error</th>
<th>Percent of Variance Accounted For</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.66</td>
<td>0.34</td>
<td>.1898</td>
<td>94.3</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>0.00</td>
<td>.3848</td>
<td>86.2</td>
</tr>
<tr>
<td>3</td>
<td>0.30</td>
<td>0.70</td>
<td>.0633</td>
<td>97.8</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>1.00</td>
<td>.1965</td>
<td>92.6</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>0.00</td>
<td>1.1814</td>
<td>62.9</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>1.00</td>
<td>.0301</td>
<td>99.6</td>
</tr>
<tr>
<td>7</td>
<td>0.64</td>
<td>0.36</td>
<td>.1148</td>
<td>94.7</td>
</tr>
</tbody>
</table>

The data were analyzed a second time for Model II. Parameter a' had been set to zero, in accord with the assumption that Ss would use only the voicing feature in making their judgment. Parameter b, the weight of the interaction term in Model II, was allowed to vary from 0.0 to 1.0 in increments of .02. The squared error between predicted and observed identification functions was also computed. The values for each S that resulted in minimum squared error are shown.
Figure 3: Identification functions for a representative S. Part A is for the /ba/ to /da/ series, part B the /ba/ to /pa/ series, part C the /ba/ to /ta/ series with two response alternatives, and part D is for the /ba/ to /ta/ series with four response alternatives.
in Table 2. The proportion of the variance accounted for by the predicted function was computed and is shown in Table 2. The proportion of the variance accounted for by Model II is greater than or equal to that accounted for by Model I for every $S$. The overall mean proportion of variance accounted for was 92.9 percent in Model II.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Constants</th>
<th>Minimum Squared Error</th>
<th>Percent of Variance Accounted For</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>1.00</td>
<td>.1831</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.00</td>
<td>.1790</td>
</tr>
<tr>
<td>3</td>
<td>0.00</td>
<td>1.00 0.30</td>
<td>.0570</td>
</tr>
<tr>
<td>4</td>
<td>0.00</td>
<td>1.00</td>
<td>.1064</td>
</tr>
<tr>
<td>5</td>
<td>0.00</td>
<td>1.00</td>
<td>.6570</td>
</tr>
<tr>
<td>6</td>
<td>0.00</td>
<td>1.00 0.00</td>
<td>.0301</td>
</tr>
<tr>
<td>7</td>
<td>0.00</td>
<td>1.00 0.64</td>
<td>.1154</td>
</tr>
</tbody>
</table>

Both Model I and Model II predict the identification probabilities along the bidimensional speech series reasonably well. However, predictions from Model II, the interaction model, fit the observed probabilities somewhat better than predictions from the additive model. There was an increase in the proportion of variance accounted for in four out of the seven $S$s with Model II. For three of the $S$s the variance accounted for by Model II remained the same as in Model I, although the parameter values changed. In fact, the three $S$s with the highest proportion of variance accounted for in Model I are the three $S$s for whom Model II shows no gain.

We suggested earlier that identification of the bidimensional series /ba/ to /ta/ might be based on the use of only one feature—voicing—since $S$s were constrained to only two response categories. Parameter $a'$ in Model II was set at zero on the assumption that the place feature is based entirely on the voicing feature and would not contribute directly to the response decision. The strength of the interaction model, Model II, can be tested by letting parameter $a'$ vary as in equation 8. Accordingly, when the squared error between the predicted and observed probabilities was obtained by equation 8, $a'$ was estimated to be zero for every $S$. The estimates of parameter $b$ were identical to those obtained with equation 9 where $a'$ was previously set to zero. This suggests that our original assumption was correct. $S$s apparently relied more on the voicing feature than the place feature in the two category bidimensional series.

The extent to which place information enters into the voicing decision for each $S$ is reflected in parameter $b$ from Model II (equation 9). This parameter is greater than zero for all $S$s except one, indicating that place information does affect the voicing decision, although only in terms of an interaction. Although the fit of the additive model (Model I) is good, the better fit of the interactive
model (Model II) and the generally nonzero estimates of the interaction term support the notion that the phonetic features of place and voicing in stop consonants are not combined independently in stage 3.

Probabilities for the second identification function generated by Ss for the /ba/ to /ta/ series with four response alternatives were also computed. Although this condition was included in the experiment almost as an afterthought, the results were not only surprising but consistent among Ss. The identification function for the same representative S as before in this condition is shown in Figure 3D. The high probability of a P response for stimulus 7 and the distribution of P responses around this mode is of special interest. If a S were responding P at random, the Pr[P] in this series should be .25 for all stimuli instead of approximately zero everywhere except for a few stimuli. This same pattern of P responses was found for all Ss tested. The peak probability of a P and the stimulus at which it occurred are shown in Table 3. When the data for each S is broken down by tape presentation, the same results are observed (see Table 3).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.90</td>
<td>1.00</td>
<td>.80</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>.85</td>
<td>.90</td>
<td>.80</td>
<td>6, 7</td>
</tr>
<tr>
<td>4</td>
<td>.80</td>
<td>.70</td>
<td>.90</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>.80</td>
<td>.90</td>
<td>.70</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>.85</td>
<td>1.00</td>
<td>.70</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>.65</td>
<td>.90</td>
<td>.40</td>
<td>7</td>
</tr>
</tbody>
</table>

In contrast, the Pr[D] was much lower for all Ss except one, and showed greater variability when subject to split-half analysis. These data are shown in Table 4. One S reported only a single /da/ in 220 test trials.

It would appear that the occurrence of /da/ identifications in the /ba/ to /ta/ series with four response categories may be randomly distributed. On the other hand, the occurrence of /pa/ identifications is highly consistent both within and across Ss and the peak probability is never less than .65.

If the phonetic features of place and voicing combined separately and additively in stage 3 as Model I would predict, the identification functions for this second series should resemble the data for the first /ba/ to /ta/ series. This did not occur, as shown in Figure 3D. Ss showed consistent use of the /pa/ response in the second /ba/ to /ta/ series at levels well above chance expectation. The peak Pr[P] in this second bidimensional series occurred at a stimulus whose place value generally corresponded to a high Pr[D] in the /ba/ to /da/ (place) series (see Table 5). Similarly, the peak Pr[P] stimulus in the bidimensional
TABLE 4: Peak Pr[D] in the second /ba/-/ta/ series.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Peak Pr[D]</th>
<th>Split-Half Pr[D] 1</th>
<th>Split-Half Pr[D] 2</th>
<th>Stimulus Where Peak Occurs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.10</td>
<td>0.00</td>
<td>.20</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>.80</td>
<td>.80</td>
<td>.80</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>.45</td>
<td>.60</td>
<td>.30</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>.35</td>
<td>.60</td>
<td>.10</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>.50</td>
<td>1.00</td>
<td>0.00</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>.05</td>
<td>.10</td>
<td>0.00</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>.30</td>
<td>.20</td>
<td>.40</td>
<td>5</td>
</tr>
</tbody>
</table>

TABLE 5: Peak Pr[P] in the second /ba/-/ta/ series; Pr[D] in the /ba/-/da/ series; and Pr[P] in the /ba/-/pa/ series to the corresponding stimulus.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Peak Pr[P]</th>
<th>Corresponding Pr[D]</th>
<th>Corresponding Pr[P]</th>
<th>Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.90</td>
<td>.267</td>
<td>.967</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>.800</td>
<td>.967</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>.85</td>
<td>.500</td>
<td>.967</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>.80</td>
<td>.900</td>
<td>.933</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>.80</td>
<td>.933</td>
<td>.933</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>.85</td>
<td>.733</td>
<td>.833</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>.65</td>
<td>.867</td>
<td>1.000</td>
<td>7</td>
</tr>
</tbody>
</table>

series corresponds to a stimulus in the /ba/ to /pa/ (voicing) series which exhibits a high Pr[P] (see Table 5). A model that assumes separate, additive weighting of the features, such as Model I, would predict that the stimulus where the peak Pr[P] occurs in the second bidimensional condition would be categorized as /ta/ and not /pa/.

A model to fit these four-response data was constructed based on equation 8. However, since Model II did not fit the four-response data very well, even when parameter a' was allowed to vary, another term was added in which the place feature is dependent on the voicing feature. This model now reflects an interdependence of these two features on each other. This model, summarized in equation 10 below, has two parameters to be estimated:

\[
Pr[T] = a'Pr[D] - b'(1-Pr[P]) Pr[D] + (1-a') Pr[P] + (1-b') (1-Pr[D]) Pr[P]
\]

(10)
Figure 4: Observed identification function for a representative S in the four- response /ba/ to /ta/ series (part A) and the predicted function for the same S using equation 10 (part B).
Equation 10 generally failed to predict the magnitude of the Pr[P] in the four-response condition, although equation 10 did generally predict a peak at stimulus seven. The fit of equation 10 to one S's data is shown in Figure 4. The obtained identification function is shown in panel A; the predicted function derived by equation 10 is shown in panel B. The response data for all conditions for this same S were shown previously in Figure 3.

The failure of equation 10 to predict accurately the entire set of probabilities for the second /ba/ to /ta/ series may be attributed to two possible factors. First, Ss' identification functions for the two component series (/ba/ to /da/ and /ba/ to /pa/) were somewhat noisy. Second, processing in stages 1 and 2 of Figure 2 may not be independent as we have assumed. Any nonindependence of processing, especially in stage 2 where the phonetic features are extracted, would affect the assumptions made in deriving Model I and Model II.

In summary, an additive model which assumes independence in the processing of phonetic features cannot account for the identification functions when the acoustic cues underlying place and voicing in stop consonants are varied systematically. Rather, it appears that an interaction model handles the data much better and provides additional support for the evidence previously reported by Lisker and Abramson (1964, 1967) and Haggard (1970) with different experimental procedures. The perception of an acoustic cue underlying a particular phonetic feature (e.g., place or voicing) may not be invariant with changes in the acoustic cues underlying other phonetic features. This conclusion is scarcely surprising since covariances in the acoustic cues derive directly from production constraints, and is added evidence of the close link between speech perception and production.

CONCLUSION

An additive model which assumes independence of processing at all stages did a creditable job in predicting the response probabilities along a bidimensional series of synthetic stop consonants when Ss were constrained to two responses. However, a model that does not assume additive processing (i.e., nonindependence) in the stage where phonetic features are combined does even better than the independence (i.e., additive) model. When Ss are given four responses from which to choose, the additive model fails completely. In contrast, the nonadditive model, while not yielding an excellent fit, does predict occurrence of /pa/ identifications on the /ba/ to /ta/ series. Based on these perceptual data with synthetic speech stimuli, we conclude that phonetic features in stop consonants are not combined independently to form phonetic segments.

REFERENCES


Summerfield, A. Q. and M. P. Haggard. (1972) Perception of stop voicing. Speech Perception (Queen's University of Belfast, Northern Ireland) 2.