Reaction Times to Comparisons Within and Across Phonetic Categories: Evidence for Auditory and Phonetic Levels of Processing

David B. Pisoni* and Jeffrey Tash*

Same-different reaction times (RTs) were obtained in a Posner-type matching task to pairs of synthetic speech sounds ranging perceptually from /ba/ through /pa/. Listeners were required to respond "same" if both stimuli in a pair were the same phonetic segments (i.e., /ba/-/ba/ or /pa/-/pa/) or "different" if both stimuli were different phonetic segments (i.e., /ba/-/pa/ or /pa/-/ba/). RT for "same" matches was faster to pairs of acoustically identical stimuli (A-A) than to pairs of acoustically different stimuli (A-a) belonging to the same phonetic category. RT for "different" responses was slower for a two-step difference across the phonetic boundary than for a four-step or six-step difference. These results provide evidence for distinct auditory and phonetic levels of processing in speech perception. Low-level acoustic information about stop consonants may be available to listeners but this is dependent on the level of processing accessed by the particular information processing task employed.

It is now well established that when listening in the speech mode a subject can identify the phonetic category of a sound but cannot discriminate between acoustically different sounds selected from within the same phonetic category (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Liberman, 1970; Pisoni, 1971, 1973). This phenomenon, known as "categorical perception," appears to be unique to certain classes of speech sounds. In the idealized case, two speech sounds can be discriminated only to the extent that they can be identified as different on an absolute basis (Studdert-Kennedy, Liberman, Harris, and Cooper, 1970). This contrasts with other kinds of auditory perception where discrimination is better than absolute identification (Pollack, 1952, 1953).

*An earlier report of these findings was presented at the 85th meeting of the Acoustical Society of America, Boston, Mass., 11 April 1973.

+Indiana University, Bloomington.

Acknowledgment: This research was supported by a grant from NICHD to Haskins Laboratories and a PHS grant (S05 RR 7031) to Indiana University. We are very grateful to Professor Lloyd Peterson for the use of his computer and to R. M. Shiffrin, M. Studdert-Kennedy, and I. Pollack for their interest and advice on this project.

[HASKINS LABORATORIES: Status Report on Speech Research SR-34 (1973)]
Auditory information from the earliest stages of perceptual processing of speech sounds may be lost as a consequence of phonetic categorization. Thus, acoustic information will be unavailable for use in a subsequent discrimination task (Pisoni, 1973). The complexity of speech sounds and the status they have as linguistic segments in language may force listeners to respond to these sounds in an absolute sense, transforming the sounds into more durable phonetic representations. Since the discrimination tasks employed in most speech perception experiments place a heavy load on short-term memory, it is reasonable to suppose that a phonetic representation would be stored in short-term memory in preference to the auditory transform of the complex acoustic signal. Accordingly, the observed "categorical" discrimination performance may not actually be based on the specific acoustic properties of the stimuli, but rather, on a higher, more abstract phonetic level of analysis. It is possible that information from the earliest stages of processing might be available to a listener, at least for a short period of time. However, the extent to which this relatively unencoded, low-level information can be accessed will depend on a variety of different factors including the stage or stages of perceptual analysis examined by a particular information processing task.

The present study is concerned with how listeners go from one level of perceptual analysis to another in speech perception and with what type of information may remain of previous levels of analysis. Specifically, we were concerned with determining whether listeners could respond to acoustic differences between categorically perceived speech sounds or whether they can only process these sounds on an abstract phonetic basis. The procedure used to investigate this problem was the reaction time (RT) matching paradigm developed by Posner and his associates (Posner and Mitchell, 1967; Posner, 1969; Posner, Boies, Eichelman, and Taylor, 1969). This procedure provides an opportunity to examine the levels of analysis at which comparisons are made by measuring the processing time required for different types of comparisons.

Thus, when a listener is asked to determine whether two speech sounds are the "same" or "different," the time to arrive at a decision may reflect the level of perceptual processing and in turn the type of information required for a comparison. Some speech sounds may be compared directly, based on their acoustical properties, while other stimuli may require a process of abstraction where invariant features must first be identified before being compared (Posner and Mitchell, 1967; Posner, 1969). Classifying two acoustically different speech sounds as the "same" may be considered to involve matching abstracted phonetic features at a higher level of perceptual analysis than classifying two acoustically identical stimuli as the "same." The latter comparison could be based on an earlier stage of analysis involving only the low-level acoustic properties of the stimuli.

Figure 1 shows a flowchart of a model of the stages of analysis involved in this type of classification task. This model is adapted from Posner and Mitchell's (1967) work on letter classification.

On every trial a listener is presented with a pair of stimuli and is required to determine whether the members of the pair are the "same" or "different." Three types of stimulus pairs are shown at the top of the figure, A-A, A-a, and A-B. The A-A pairs represent acoustically identical pairs of stimuli. The A-a pairs represent acoustically different stimuli selected from within a
FLOWCHART OF CLASSIFICATION TASK

Figure 1: Model of the stages of analysis involved in the "same" - "different" classification task.
particular phonetic category. [In Posner's (1969) terminology, these would be pairs of physically different stimuli with the same "name" code.] Finally, the A-B pairs represent stimuli selected from different phonetic categories. These are acoustically different and have different names.

Depending on the particular type of stimulus pair, various predictions can be made about the relative amount of time required for "same" matches and "different" matches. For example, if low-level acoustic information can be accessed for a comparison, reaction time should be faster for a "same" response when the input pairs are acoustically identical (e.g., A-A) than when they are acoustically different but phonetically the same (e.g., A-a). This should be true if the acoustically identical pairs (e.g., A-A) could be matched as "same" at an earlier stage of analysis than the acoustically different pairs (e.g., A-a). The acoustically different pairs would require an additional stage of analysis for a "same" response. However, if only an abstract phonetic representation is used in the comparison, reaction times for a "same" match to these two types of pairs should be identical. Under this assumption, a similar set of predictions can also be made for the "different" matches. If distinct auditory and phonetic levels of processing exist, pairs of stimuli with large physical differences should be matched as "different" faster than pairs of stimuli with smaller physical differences. If only an abstract phonetic representation is available, reaction time for "different" matches should be equivalent, regardless of the magnitude of the physical differences between pairs of stimuli.

**METHOD**

**Subjects**

The listeners were nine paid volunteers, all of whom were either graduate students or staff members associated with the Mathematical Psychology Program at Indiana University. The Ss were right-handed native speakers of English and reported no history of a hearing disorder or speech impediment. Ss were paid for their services at the rate of $1.50 per hour. All Ss had had some previous experience with synthetic speech stimuli, although they were naive to the exact purposes of the present experiment.

**Stimuli**

A set of bilabial stop consonant-vowel (CV) stimuli were synthesized on the parallel resonance synthesizer at Haskins Laboratories. The basic set of stimuli consisted of seven three-formant syllables 300 msec in duration. The stimuli varied in 10-msec steps along the voice onset time (VOT) continuum from 0 through +60 msec, which distinguishes /ba/ and /pa/. VOT has been defined as the interval between the release of the articulators and the onset of laryngeal pulsing or voicing (Lisker and Abramson, 1964). Synthesizer control parameter values for these stimuli were similar to those employed by Lisker and Abramson (1970) in their cross-language experiments. The final 250 msec of the CV syllable was a steady-state vowel appropriate for an English /a/. The frequencies of the first three formants were fixed at 769, 1,232, and 2,525 Hz respectively. During the initial 50-msec transitional period, the first three formants moved upward toward the steady-state frequencies of the vowel. For successive stimuli in the set, the delay in the rise of F1 to full amplitude (i.e., the degree of F1 "cutback") and in the switch of the excitation source
from hiss (aperiodic) to buzz (periodic) was increased by 10 msec. Simultaneous changes in amplitude in the lower frequency region and type of excitation source have been shown to characterize the voicing and aspiration differences between /b/ and /p/ in English (Lisker and Abramson, 1967).

Experimental Materials

All stimuli were digitized and their wave forms stored on the Pulse Code Modulation System at Haskins Laboratories (Cooper and Mattingly, 1969). Two types of audio tapes were prepared under computer control: an identification test and a matching test. A 1,000 Hz tone of 100 msec duration was recorded 500 msec before the onset of each trial. This tone served as a warning signal for the Ss and was also used to trigger a computer interrupt which initiated timing response latency.

Two different 140-trial identification tests were prepared. Each test contained 20 different randomizations of the seven stimuli. Stimuli were recorded singly with a 3-sec interval between presentations. Each stimulus occurred equally often within each half of the tape.

Four different "same" - "different" matching tests were constructed. Each test tape contained 48 pairs of stimuli. Half of all the trials consisted of within-category pairs requiring a "same" response while the other half consisted of across-category pairs requiring a "different" response. Figure 2 shows the arrangement of the stimulus conditions employed in the present experiment.

Within-category pairs were either physically identical (A-A) or physically different (A-a). A-A trials consisted of stimuli 1, 3, 5, and 7, each paired with itself (i.e., 1-1, 3-3, 5-5, 7-7). The A-a trials, which were separated by two steps along the continuum or +20 msec VOT, consisted of the stimulus pairs 1-3, 3-1, 5-7, and 7-5.

Across-category pairs (A-B), which were always physically different, were separated by two, four, or six steps along the continuum. These comparisons represented differences of +20, +40, or +60 msec VOT respectively.

Each of the eight within-category comparisons appeared three times within a block of 48 trials, whereas each of the six between-category comparisons appeared four times. The interstimulus interval (ISI) between members of a pair was held constant at 250 msec. Successive trials were separated by 4 sec.

Procedure

The experimental tapes were reproduced on an Ampex AG-500 two-track tape recorder and were presented diotically through Telephonics (TDH-39) matched and calibrated headphones. The gain of the tape recorder was adjusted to give a voltage across the earphones equivalent to 70 db SPL re 0.0002 dynes/cm² for a 1,000 Hz calibration tone. Measurements were made on a Hewlett Packard VTVM (Model 400) before the presentation of each experimental tape. Ss were run individually in a small experimental room. All responses and reaction times were recorded automatically under the control of a PDP-8 computer located with the tape recorder in an adjacent room.
Figure 2: Description of the stimulus conditions employed in the matching task. Pairs of stimuli requiring a "same" response are selected from within a phonetic category; pairs of stimuli requiring a "different" response are selected from across phonetic categories.
The instructions for the identification test were similar to those used in previous speech perception experiments. Ss were required to identify each stimulus as either /ba/ or /pa/ and to respond as rapidly as possible. The Ss responded to each stimulus by pressing one of two labeled telegraph keys. For a given S, one hand was always used for a /ba/ response while the other hand was used for a /pa/ response. The two keys were counterbalanced for hands across Ss.

For the matching task, Ss were told that they would hear a pair of stimuli on every trial and their task was to decide whether the two stimuli were the "same" or "different" phonetic segments. The type of instructions employed here is similar to the "name match" instructions employed by Posner and Mitchell (1967). Ss were told that half of all the pairs were the same (e.g., /ba/-/ba/ or /pa/-/pa/) and half of the pairs were different (e.g., /ba/-/pa/ or /pa/-/ba/). Ss were encouraged to respond as rapidly as possible. As in the identification task, Ss responded to each pair by pressing one of two telegraph keys, labeled "same" and "different." The response keys were also counterbalanced for hands across Ss.

Ss were tested for an hour a day on two consecutive days. Each session began with a 140-item identification test which was followed after a short break by two 48-trial matching tests. Since the first day served as a practice session, only the identification and matching data from the second session will be considered in the remainder of this report.

RESULTS AND DISCUSSION

Identification Task

The average identification function is shown in Figure 3 along with the mean RT for identification. Each point represents the mean of 180 responses over the nine Ss.

The filled squares and open circles show percent /ba/ or /pa/ response respectively to each of the seven stimuli in the continuum. The filled triangles represent the corresponding latency of identification response to each stimulus. Examination of Figure 3 indicates that the identification function is quite consistent. Ss partitioned the stimulus continuum into two discrete phonetic segments. The phonetic boundary or cross-over point in identification is at about +30 msec VOT which is consistent with previous findings (Lisiker and Abramson, 1967).

Inspection of the RT function during identification shows that Ss are slowest for stimulus 4 which is at the phonetic boundary and fastest for the other stimuli which are within phonetic categories. These results are also consistent with the findings reported by other investigators who have studied reaction time in the identification of synthetic speech sounds (Studdert-Kennedy, Liberman, and Stevens, 1963). Reaction time is a positive function of uncertainty, increasing at the phonetic boundary where identification is least consistent and decreasing where identification is most consistent. In anticipation of the discrimination tests, it is noted that identification time is slowest for the stimulus region where discrimination is best.
Figure 3: Average identification function for the voice onset time continuum with mean RT during identification.
Matching Task

The major results of the "same" - "different" classification task are shown in Figure 4.

The mean RT for each of the two types of "same" trials (A-A, A-a) is based on a total 216 judgments, while the mean RT for each of the three types of "different" pairs is based on 144 judgments averaged over nine Ss.

An examination of the "same" responses reveals that subjects are faster for pairs of acoustically identical stimuli (e.g., A-A) than for pairs of acoustically different stimuli (e.g., A-a) which have been selected from within a phonetic category. The 41 msec difference between these two conditions is highly significant (P<.01) by a correlated t-test, t(8) = 3.20 (one-tailed). This result is consistent with the model described earlier. Ss can access low-level acoustic information even though they have categorized these pairs of stimuli as the "same" phonetic segments. Thus, "same" matches to acoustically identical speech sounds are presumably based on an earlier stage of perceptual analysis than "same" matches to acoustically different speech sounds. In the latter case, the "same" response is based on a comparison of the phonetic features of each stimulus which must have been extracted before a match could have been made. It may be assumed that the abstraction of phonetic features from the acoustic signal requires an additional amount of processing time. This is presumably responsible for the difference in RT between the two within-category conditions.

The present findings, based on "same" responses to within-phonetic category comparisons indicate that even perception of stop consonants may not be entirely categorical, as previously supposed. Rather, the degree of categorical perception will depend upon the extent to which low-level acoustic information can be utilized within the experimental task. Since acoustic information not only decays rapidly over time but also is highly vulnerable to various types of interfering stimuli, the specific discrimination procedure may be crucial in determining the relative roles of acoustic and phonetic information in speech sound discrimination. For example, the ABX procedure may force the listener to rely almost entirely on phonetic information in discrimination because of the arrangement of stimuli in this procedure.

One additional point should be emphasized here concerning the within-phonetic category comparisons. It could be argued that, in the present experiment, Ss were responding to these stimuli as isolated acoustic events rather than as speech sounds. If so, the proportion of "same" responses should have been quite different for A-a pairs and A-A pairs. In fact, P('same'|A-a) and P('same'|A-A) were almost identical, suggesting that Ss were responding to these stimuli as speech sounds.

The mean RTs for "different" responses to the three types of across-category pairs are also shown in Figure 4. An examination of these RTs provides additional support for the argument that Ss can employ relatively low-level acoustic information in the comparison process. RT for "different" matches is slower for a two-step difference than for a four-step or six-step difference across the phonetic boundary. Both differences are significant (p<.005) by correlated t-tests, t(8) = 4.95; t(8) = 4.82, respectively. These findings
Figure 4: Mean RT for "same" and "different" responses to within- and across-category comparisons. The number of steps between pair members reflects the magnitude of the acoustic difference in voice onset time (VOT).
suggest that "different" matches may not be based solely on an abstract phonetic representation. Rather, a "different" response to pairs of stimuli across category boundaries may also be based on low-level acoustic information at an earlier stage of perceptual analysis. Pairs of stimuli which are separated by large physical differences in VOT, such as 1-7 and 2-6, can be differentiated solely on the basis of their acoustic dissimilarity. Stimulus pairs separated by small differences in VOT, such as 1-3 and 3-5, cannot be differentiated on the basis of their acoustic similarity and an additional stage of analysis is required. Since an initial decision cannot be made reliably on the basis of acoustic information alone, the "different" decision for pair 3-5 must, therefore, be based on a comparison of the phonetic features of the two stimuli.

In summary, the results suggest that low-level acoustical information about a speech stimulus may be available to listeners along with a more abstract phonetic representation, even in the case of stop consonants. Presumably the extent to which low-level information can be accessed will depend not only on the particular level of perceptual analysis examined but also on the type of information processing task employed.

The results of this experiment argue for a diversity of experimental tasks in the study of speech sound perception. On the basis of the distribution of responses alone, we might conclude that only a categorical or phonetic analysis is available for stop consonants. The addition of the RT task reveals another level of analysis. A view of speech sound perception entailing a series of interrelated stages of analysis could serve as the framework for determining quantitatively the ways in which speech perception may involve specialized mechanisms for perceptual analysis. Moreover, such an approach may help to determine the ways in which various speech perception phenomena may conform to more general perceptual processes.

REFERENCES


