RESEARCH ON SPEECH PERCEPTION

Progress Report No. 5

October 1978 - August 1979

David B. Pisoni

Principal Investigator

Department of Psychology

Indiana University

Bloomington, Indiana 47405

Supported by:

Department of Health, Education, and Welfare
U.S. Public Health Service

National Institute of Mental Health
Grant No. MH-24027-05

and

National Institutes of Health
Grant No. NS-12179-04

CONTENTS

INTRODUCTION

This is the fifth annual progress report of research activities on speech perception, analysis and synthesis conducted in the Department of Psychology at Indiana University. As with our previous progress reports, our main goal has been to summarize our various research activities over the past year and make them available to interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief progress reports on the status of on-going research projects in the laboratory. We also have included new information on instrumentation developments and software support when we think this information would be of interest or help to other colleagues.

We decided to issue progress reports of our research activities several years ago primarily because of the lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis and synthesis and, therefore, we would be most grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Department of Psychology
Indiana University
Bloomington, Indiana 47405
U.S.A.

I. EXTENDED MANUSCRIPTS

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 5 (1978-79) Indiana University]

SOME MEASURES OF INTELLIGIBILITY AND COMPREHENSION*

David B. Pisoni

Research Laboratory of Electronics

Massachusetts Institute of Technolology

Cambridge, Massachusetts 02139

*This chapter was prepared for a special summer session course entitled Conversion of Unrestricted English Text to Speech held at MIT on June 25-29, 1979. The final version of the chapter is to appear in a forthcoming research monograph edited by Jonathan Allen that describes the entire MIT Text-to-Speech system.

CHAPTER 13

SOME MEASURES OF INTELLIGIBILITY AND COMPREHENSION*

David B. Pisoni

13.1 OVERVIEW

As the ten year effort to build an unrestricted text-to-speech system at MIT drew to a close, it seemed appropriate to conduct a preliminary evaluation of the quality of the speech output with a relatively large group of naive listeners. The results of such an evaluation would no doubt prove useful in first establishing a benchmark level of performance for comparative purposes as well as uncovering any problems in the current version of the system that might not have been detected earlier. In addition to obtaining measures of

*Copyright (c) 1979 by David B. Pisoni

intelligibility of the speech output produced by the text-to-speech system, we were also interested in finding out how well naive listeners could comprehend continuous text produced by the system. This was thought to be an important aspect of the evaluation of the text-to-speech system as a whole since a version of the current system might eventually be implemented as a device used for computer aided instruction or as a functional reading machine for the blind (Allen, 1973). Both of these applications are now well within the realm of the available technology (Allen, Hunnicutt, Carlson & Granstrom, 1979).

In carrying out the evaluation of the system, we patterned several aspects of the testing after earlier work already completed on the evaluation of the Haskins Laboratories reading machine project so that some initial comparisons could be drawn between the two systems (Nye & Gaitenby, 1973, 1974). However, we also added several other tests to the evaluation to gain additional information about word recognition in normal sentential contexts and listening comprehension for a relatively wide range of narrative passages of continuous text. Data were also collected on reading comprehension for the same set of materials, to permit direct comparison between the two input modalities. Traditional measures of listening or reading comprehension have not typically been obtained in previous evaluations of the quality of synthetic speech output and therefore we felt that some preliminary data would be quite useful before the major components of the present system were implemented as a workable text-to-speech device in an applied context.

In planning the current evaluation project, we also wanted to obtain information about several different aspects of the total system and their contribution to intelligibility and comprehension of speech. To meet this goal, a number of different tests were selected to provide information about: (1) phoneme recognition, (2) word recognition in sentences and (3) listening comprehension. It was assumed that the results of these three tests together would provide qualitative and quantitative information sufficient to identify any major problems in the operation of the total system at the time of testing in early May of 1979. The results of these three types of tests would also provide us with much more detailed information about the relative contribution of several of the individual components of the system and their potential interaction.

In carrying out these evaluation tests, we collected a total of 27,128 responses from some 160 naive listeners. A total of 45 minutes of synthetic speech was generated in fully automatic text-to-speech mode. No system errors were corrected at this time and no total system crashes were encountered during the generation of the test materials used in the evaluation.

13.2 PHONEME RECOGNITION

After initial discussions, we decided to use the Modified Rhyme Test to measure the intelligibility of the speech produced by the system. This test was originally developed by Fairbanks (1958) and then later modified by House, Williams, Hecker and Kryter (1965). This test was chosen primarily because it is reliable, shows little effect of learning and is easy to administer to untrained and relatively naive listeners. It also uses standard orthographic responses, thereby eliminating problems associated with phonetic notation. Moreover, extensive data have already been collected with natural speech as well as synthetic speech produced by the Haskins speech synthesizer (Nye and Gaitenby, 1973) therefore permitting us to make several direct comparisons of the acoustic-phonetic output of the two text-to-speech systems under somewhat comparable testing conditions.

Method

Subjects. Seventy-two naive undergraduate students at Indiana University in Bloomington served as paid listeners in this study. They were all recruited by means of an advertisement in the student newspaper and reported no history of a hearing or speech disorder at the time of testing. The subjects were all right-handed native speakers of English.

Stimuli. Six lists of 50 monosyllabic words were prepared on the MIT text-to-speech system. The lists were recorded on audio tape via a Revox Model B77 tape recorder at 7 1/2 ips with a 3.0 second pause between successive items. Approximately half of the items in a given test list differed in the initial consonant while the remaining half differed in the final consonant.

Procedure. The seventy-two subjects were divided up into twelve independent groups containing six subjects each for testing. Two groups of subjects were assigned to each of the six original test lists. Subjects were told that this was a test dealing with isolated word recogntion and that they were to indicate which word out of six possible alternatives was the one they heard on each trial. Forced-choice response forms were provided to subjects to record their judgements. Subjects were encouraged to guess if they were not sure but to respond on each trial. No feedback was provided to subjects during the course of testing. Subjects were, however, explicitly informed that the test items were generated on a computer and that the experiment was designed to evaluate the intelligibility of the synthetic speech. An example of the test format is provided in Appendix A.

Testing was carried out in a small experimental room in the Speech Perception Laboratory in the Department of Psychology at Indiana University. This room is equipped with six individual cubicles. The audio tapes were reproduced on an Ampex AG-500 taperecorder and presented to subjects via TDH-39 matched and calibrated headphones at a comfortable listening level of about 80 dB SPL peak reading on a VTVM. A low-level (60 dB) broad band (0-10 kHz) white noise source (Grason Stadler Model 1724) was also mixed with the speech to mask tape hiss, some non-stationery computer generated background noise picked up during the recording at MIT, and any ambient noise in the local environment during testing.

Results and Discussion

Although the Modified Rhyme Test employed real words, our interest was focused on the phoneme errors and resulting percepual confusions. Overall performance on the test was very good with a total error rate, averaged across both initial and final syllable positions, of only 6.9 %. Performance was somewhat better for consonants in initial position (4.6 % errors) than final position (9.3 % errors).

The distribution of all errors across various manner classes is shown graphically in Figure 1 for initial and final syllable position separately.

Since the consonants comprising the various manner classes occured with unequal frequencies in the Modified Rhyme Test, the observed error rates in the data may not be representative estimates of the intelligibility of the same phonemes in continuous speech. Nevertheless, performance is generally excellent across almost all manner categories except for the nasals in final position which showed an error rate of 27.6%. It should also be noted that while consonants in initial position were identified better than the same ones in final position, the relative distribution of the errors across syllable positions is not comparable as shown in Figure 2 below.

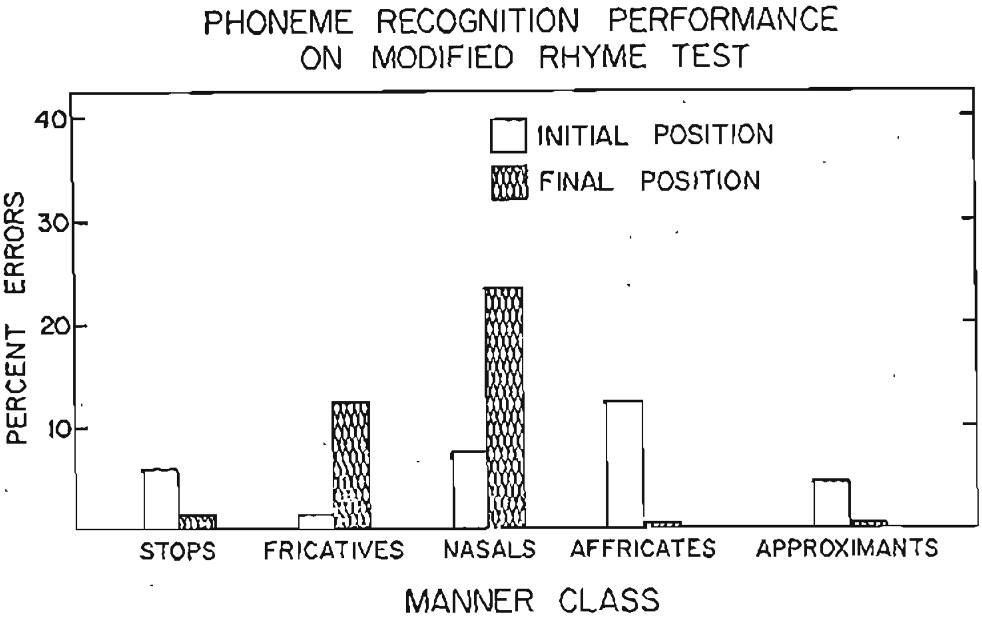


Figure 1. Average percent errors across various manner classes.

Figure 2 provides a detailed breakdown of the errors and the resulting confusions for consonants in initial and final position. Each bar in the figure shows the total percent errors for a particular phoneme and the rank order of the most frequent confusions.

In examining these data, it should be kept in mind that the error rates which make up the data shown in these two panels are quite low to begin with. Recall that the total percent errors was only 4.6 % in initial position and 9.3 % in final position. Inspection of this figure shows that for the most part the errors are predominantly confusions in place or manner of articulation. Errors in voicing, when they occured, were substantially lower. The fricatives /θ/ and /ð/ show very high error rates when

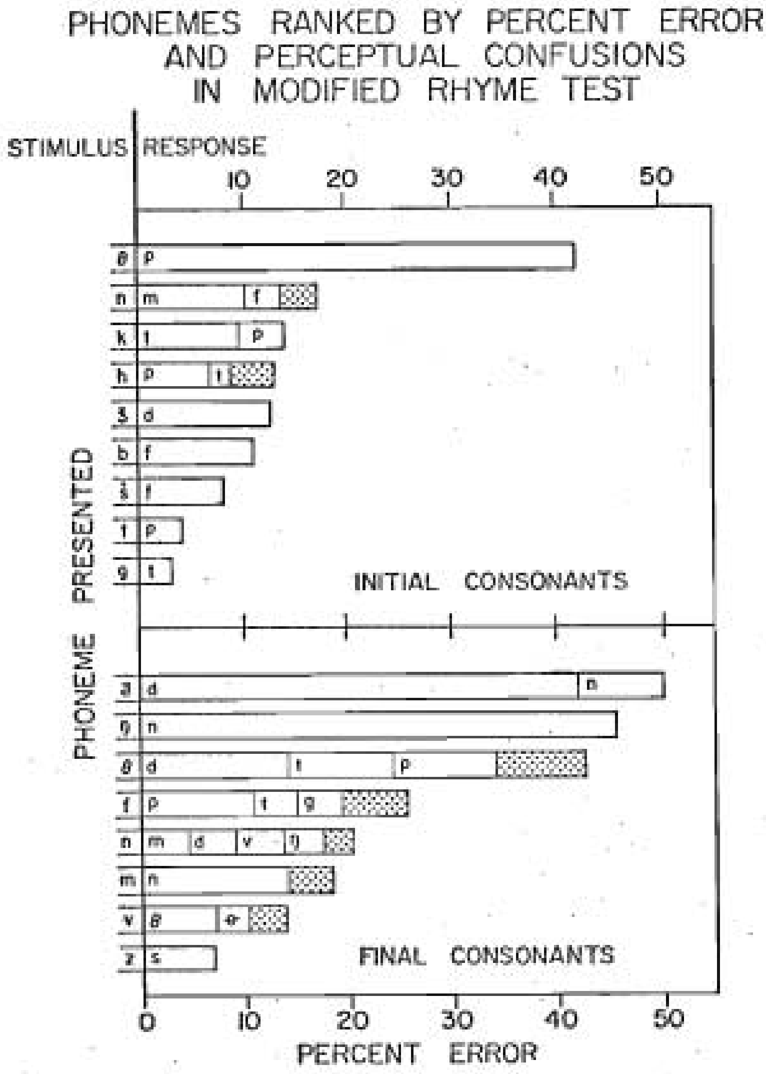


Figure 2. Distribution of errors and most frequent perceptual confusions.

considered individually although both of these phonemes occured with a relatively low frequency in the test when compared with other consonants. The presence of the background masking noise may have contributed to the low performance levels observed with these weak fricatives. As noted above, the pattern of errors is quite different for consonants in initial and final position. Such a finding is not unexpected given that different acoustic cues are used to synthesize the same phoneme in different environments.

Conclusions

For the most part, the intelligibility of the speech produced by the current version of the text-to-speech system is very high. The overall error rate of 6.9% is slightly lower than the error rate of 7.6% obtained in the earlier Haskins evaluation using the Modified Rhyme Test. The advantage of initial over final consonants observed in the present study is consistent with data obtained from natural speech by House et al. (1965) and Nye and Gaitenby (1973), although it differs slightly from the results found for the synthetic speech in the earlier Haskins evaluation. In the Haskins study, error rates for the synthetic speech in initial and final position were about the same with a very slight advantage for consonants in final position. The comparable overall error rates obtained for natural speech in the Modified Rhyme Test by House et al. and Nye and Gaitenby (1973) were 4% and 2.7% respectively.

In the earlier evaluation study, Nye and Gaitenby (1974) checked to insure that the phonemic input to the Haskins synthesizer was correct. However, no corrections of any kind were made by hand in generating the present materials either from entries in the morph lexicon or from spelling-to-sound rules. As discussed in the final section of this chapter, several different kinds of errors were uncovered in different modules as a result of generating such a large amount of synthetic speech through the system.

Except for the high error rates observed for the nasals and fricatives in final syllable position, the synthesis of segmental information in the text-to-speech system appears to be excellent, at least as measured in a forced choice format among minimal pairs of test items. With phoneme recognition performance as high as it is--nearly close to ceiling levels--it is difficult to pick up subtle details of the error patterns that might be useful in improving the quality of the output of the phonetic component of the system at the present time. In addition, the errors that were observed in the present tests might well be reduced substantially if the listeners had more experience with the speech output produced by the system. It is well known among investigators working with synthetic speech that rather substantial improvements in intelligibility can be observed when listeners become familiar with the quality of the synthesizer. Nye and Gaitenby (1974) as well as Carlson, Granstrom and Larson (1976) have reported very sizeable learning effects in listening to synthetic speech. In the latter study, performance increased from 55% to 90% correct after the presentation of only 200 synthetic sentences over a two week period. (See also our discussion of the word recognition and comprehension results below.)

In summary, the results of the Modified Rhyme Test revealed very high levels of intelligibility of the speech output from the system using naive listeners as subjects. While the overall level of performance is somewhat lower than in previous studies employing natural speech, the level of performance for recognition of segmental information appears to be quite satisfactory for a wide range of text-to-speech applications at the present time.

13.3 WORD RECOGNITION IN SENTENCES

The results of the Modified Rhyme Test using isolated words indicated very high levels of intelligibility for the segmental output of the text-to-speech system. However, the Modified Rhyme Test employs a closed response set involving a forced-choice format in what may be considered a relatively low uncertainty testing situation. In the recognition and comprehension of unrestricted text, a substantially broader range of alternatives is available to the listener since the response set is open and potentially infinite in size. Moreover, the sentential context itself provides an important contribution to intelligibility of speech, a fact that has been known for many years (Miller, Heise & Lichten, 1951; Miller & Isard, 1963).

To evaluate word recognition in sentence context, we decided to obtain two quite different sets of data. One set was collected using a small number of the Harvard Psychoacoustic Sentences (Egan, 1948). These test sentences are all meaningful and contain a wide range of different syntactic constructions. In addition, the various segmental phonemes of English are represented in these sentences in accordance with their frequency of occurrence in the language. Thus, the results obtained with the Harvard sentences should provide us with a fairly good estimate of how well we might expect word recognition to proceed in sentences when both semantic and syntactic information is available to a listener. This situation could be considered comparable, in some sense, to normal listening conditions where "top-down" knowledge interacts with sensory input in the recognition and comprehension of speech (see Pisoni, 1978; Marslen-Wilson & Welsh, 1978).

We also collected word recognition data with a set of syntactically normal but semantically anomalous sentences that were developed at Haskins Laboratories by Nye and Gaitenby (1974) for use in evaluating the intelligiblity of their text-to-speech system (see also Ingeman, 1978). These test sentences permit a somewhat finer assessment of the availability and quality of "bottom-up" acoustic-phonetic information and its potential contribution to word recognition. Since the materials are all meaningless sentences, the individual words cannot be identified or predicted from knowledge of the sentential context or semantic interpretation. Thus, the results of these tests using the Haskins anomalous sentences should provide an estimate of the upper bound on the contribution of strictly phonetic information to word recognition in sentence contexts. Since the response set is also open and essentially unrestricted, we would anticipate substantially lower levels of word recognition performance on this test than on the Harvard test; in the latter test, syntactic and semantic context is readily available and can be used freely by the listener at all levels of processing the speech input. In addition, the results of the anomalous sentence test can also be compared more-or-less directly to data collected with these same test sentences by Nye and Gaitenby (1974) and Ingeman (1978). Such comparisons should prove useful in identifying similarities and possible differences in the speech output produced by the two text-to-speech systems.

Method

Subjects. Forty-four additional naive undergraduate students were recruited as paid subjects. They were drawn from the same population as the subjects used in the previous study and met the same requirements. None of these subjects had participated in the earlier study on phoneme recognition.

Stimuli. Two sets of test sentences were prepared. One set consisted of 100 Harvard Psychoacoustic Sentences. Each sentence contained five key words that were scored as a measure of word recognition. The other set consisted of 100 Haskins anomalous sentences drawn from the original list of materials developed by Nye and Gaitenby (1974). Each of these test sentences contained four key words. Two separate test lists were recorded on audio tape with a 3 second pause between successive sentences. The sentences were output at a speaking rate in excess of 180 words per minute. As before, we did not correct any pronunication errors. Examples of both types of test sentences are given in Appendix B and C.

Procedure. Twenty-one subjects received the Harvard sentences and twenty-three received the Haskins sentences. Testing was carried out in small groups of five or six subjects each under the same listening conditions described in the previous study.

Subjects in both groups were told that this study was concerned with word recognition in sentences and that their task was to write down each test sentence as they heard it in the appropriate location on their response sheets. They were told to respond on every trial and to guess if they were not sure of a word. For the Harvard sentences, the response forms were simply numbered sequentially with a continuous underlined blank space for each trial. However, since the syntactic structure of all of the Haskins sentences was identical, the response forms differed slightly: blank spaces were provided for the four key words. Determiners were printed in the appropriate locations in standard sentence frames.

The experiment was run in a self-paced format to provide subjects with sufficient time to record their responses in the appropriate space in their booklets. However, subjects were encouraged to work rapidly in writing down their responses. The experimenter operated the tape recorder on playback from within the testing room by remote control. Thus, successive sentences in the test lists were presented only after all of the subjects in a group had finished responding to the previous test sentence, and had indicated this to the experimenter. A short break was taken half way through a testing session after completion of the first 50 sentences.

Results and Discussion

The responses were scored only for correct word recognition at this time. Phonetic errors when they occurred were not considered in the present analyses although we expect to examine these in some detail at a later time. Each subject receiving the Harvard sentences provided a total of 500 responses while each subject receiving the Haskins anomalous sentences provided 400 responses to the final analysis.

Performance on the Harvard sentences was quite good with an overall mean of 93.2% correct word recognition across all 21 subjects. The scores on this test ranged from a low of 80% to a high of 97% correct recognition. Of the 6.7% errors observed, 30.3% were omissions of complete words while the remainder consisted of segmental errors involving substitutions, deletions and transpositions. In no case, however, did subjects respond with permissible non-words that could occur as potential lexical items in English.

As expected, word recognition performance on the Haskins anomalous sentences was substantially worse than the Harvard sentences, with a mean of 78.7% correct recognition averaged over all 23 subjects. The scores on this test ranged from a low of

71% correct to a high of 85% correct. Of the 21.3% errors recorded, only 11% were omissions of complete words. The difference in error patterns, particularly in terms of the number of omissions, between the two types of sentence contexts suggests a substantial difference in the subjects' perceptual strategies in the two tests. It seems quite likely that subjects used a much looser criterion for word recognition with the Haskins anomalous sentences simply because the number of permissible alternatives was substantially greater than those in the Harvard sentences. Moreover, the presence of one standard syntactic structure probably encouraged subjects to guess more often when the acoustic cues to word identification were minimal. In addition, there seemed to be evidence of semantically based intrusions in the recall data suggesting that subjects were attempting to assign an interpretation to the input signal even though they knew before hand that all of the sentences were meaningless.

As we noted earlier, substantial learning effects occur with synthetic speech. Even after an initial period of exposure, recognition performance continues to improve. Comparisons of word recognition performance in the first and second half of each of the tests indicated the presence of a reliable learning effect. For both the Harvard and Haskins sentences, performance improved on the second half of the test relative to the first half. Although the differences were small, amounting to only about 2% improvement in each case, the result was very reliable (p < .01) across subjects in both cases.

The performance levels that we obtained with the Haskins semantically anomalous sentences are very similar to those reported earlier by Nye and Gaitenby (1974) and more recently by Ingeman (1978) using the same sentences with the Haskins synthesizer and text-to-speech system. Nye and Gaitenby (1974) reported an average error rate of 22% for synthetic speech and 5% for comparable natural speech. However, Nye and Gaitenby used both naive and experienced listeners as subjects and found rather large differences in performance between the two groups, as we noted above. This result is presumably due to familiarity and practice listening to the output of the synthesizer. We suspect that if the experienced subjects were eliminated from the Nye and Gaitenby analyses, performance would be lower than the original value reported and would therefore differ somewhat more from the present findings. Nevertheless, the error rate for these anomalous sentence produced with natural speech is still lower than the corresponding synthetic versions although it is not clear at the present time how much of the difference could be due to listener familiarity with the quality of the synthetic speech.

Conclusions

The results of the two word-recognition tests indicate moderate to excellent levels of performance with naive listeners depending on the particular test format used and the type of information available to the subject. In one sense, the results of these two tests can be thought of as approximations to upper and lower bounds on the accuracy of word recognition performance with the current text-to-speech system. On the one hand, the Harvard test sentences provide some indication of how word recognition might proceed when both semantic and syntactic information is available to a listener under normal conditions. On the other hand, the Haskins anomalous sentences direct the subjects' attention specifically to the perceptual input and therefore provide a rough estimate of the quality of the acoustic-phonetic information and sentence analysis routines available for word recognition in the absence of contextual constraints. Of course, in normal listening situations, and presumably in cases where a text-to-speech system such as the present one might be implemented, the complete neutralization of such contextual effects on intelligibility would be extremely unlikely. Nevertheless, a more detailed analysis of the word recognition errors in the Haskins anomalous sentence test might provide us with additional information that could be used to modify or improve several of the modules of the system. Whether such additional improvements at these various levels of the system will actually contribute to improved intelligibility and comprehension is difficult to assess at this time since performance with meaningful sentences is already quite high to begin with as shown by the present results obtained with the Harvard sentences.

In summary, the results of tests designed to measure word recognition in two types of sentential context showed moderate-to-excellent levels of performance with synthetic speech output from the current version of the text-to-speech system. As in the previous section dealing with the evaluation of the intelligibility of isolated words, the present results, particularly with rather diverse meaningful sentences, suggest that the quality of the speech output at the present time is probably quite satisfactory for a relatively wide range of applications requiring the processing of unrestricted text. While there is room for improvement in the quality of the output from various modules of the system, as suggested by the results of the Haskins anomalous sentences, it is not apparent whether the allocation of resources to effect such changes in the system would produce any detectable differences. Differences that might be detected, if any, might well require a very restricted listening environment in which all of the higher-level syntactic and semantic information is eliminated, a situation that is unlikely to occur when the system is implemented in an applied setting. Given these results on word recognition, however, it still remains to be determined how well listeners can understand and comprehend continuous speech produced by the system, a problem we turn to in the next section of this chapter.

13.4 COMPREHENSION

Research on comprehension and understanding of spoken language has received a great deal of attention by numerous investigators in recent years. It is generally agreed that comprehension is a complex cognitive process initially involving the input and subsequent encoding of sensory information, the retrieval of previously stored knowledge from long-term memory and the subsequent interpretation, integration or assimilation of various sources of knowledge that might be available to a listener at the time. Comprehension therefore depends on a relatively large number of diverse factors, some of which are still only poorly understood at the present time. Measuring comprehension is difficult because of the interaction of many of these factors and the absence of any coherent model that is broad enough to deal with the diverse nature of language understanding.

One of the factors that obviously plays an important role in listening comprehension is the quality of the input signal expressed in terms of its overall intelligibility. But as we have seen even from the results summarized in the previous sections, additional consideration must also be given to the contribution of higher-level sources of knowledge to recognition and comprehension. In this last section, we wanted to obtain some preliminary estimate of how well listeners could comprehend continuous text that was produced by the text-to-speech system. Previous evaluations of synthetic speech output have been concerned primarily with measuring intelligibility or listener preferences with little if any concern for assessing comprehension or understanding of the content of the materials (Nye, Ingeman and Donald, 1975). Indeed, as far as we have been able to determine, no previous formal tests of the comprehension of continuous synthetic speech have ever been carried out with a relatively wide range of textual materials specifically designed to assess understanding of the content rather than form of the speech.

To accomplish this goal, we selected fifteen narrative passages and an appropriate set of test questions from several standarized adult reading comprehension tests. The passages were quite diverse, covering a wide range of topics, writing styles and vocabulary. We thought that a large number of passages would be interesting to listen to in the context of tests designed to assess comprehension and understanding. Since these test passages were selected from several different types of reading tests they also varied in difficulty and style permitting us to evaluate the contribution of all of the individual modules of the text-to-speech system in terms of one relatively gross measure.

In addition to securing measures of listening comprehension for these passages, we also collected a parallel set of data on reading comprehension of these materials from a second group of subjects. The subjects in the reading comprehension group answered the same questions after reading each passage silently as did subjects in the listening comprehension group. This condition was included in order to permit comparison between the two input modalities. It was assumed that the results of these comprehension tests would therefore provide an initial, although preliminary, benchmark against which the entire text-to-speech system could be evaluated with materials somewhat comparable to those used in the immediate future.

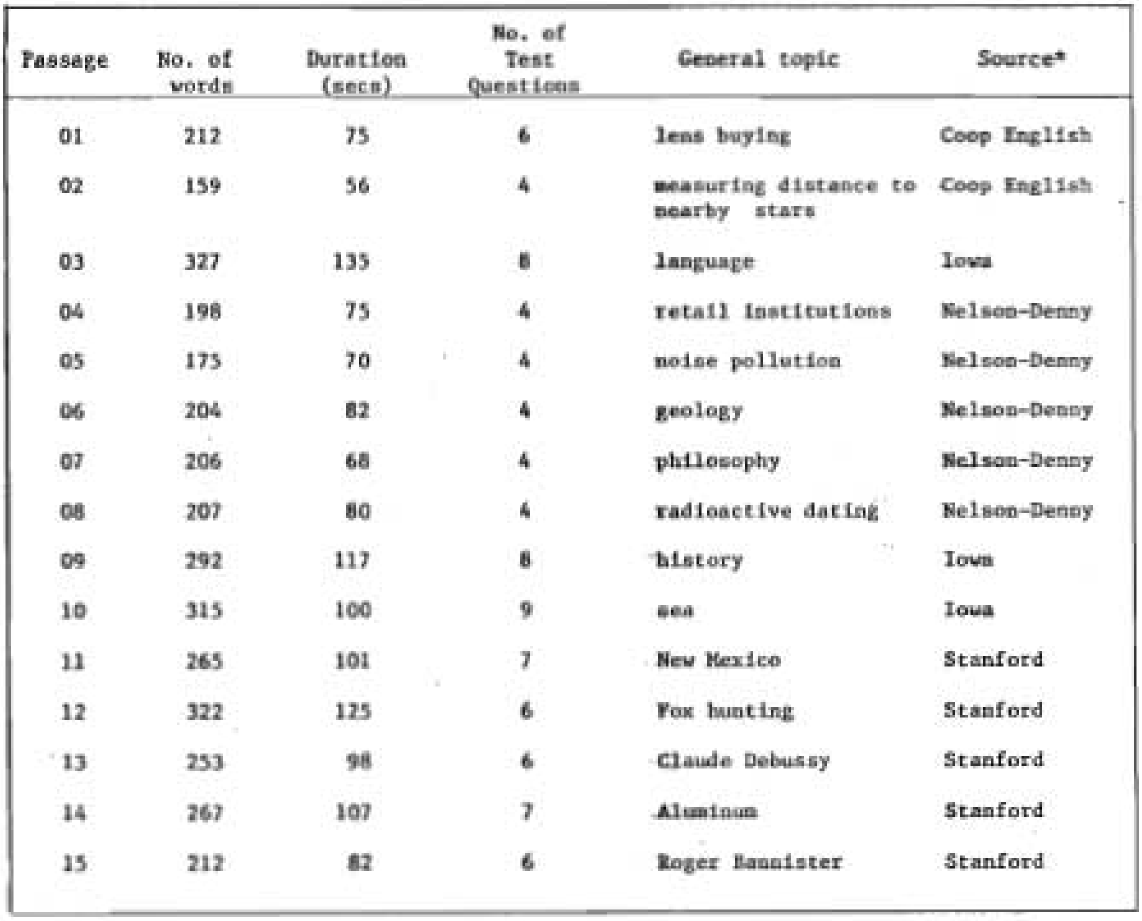
Method

Subjects. Forty-four additional naive undergraduate students were recruited as paid subjects. They were drawn from the same source as the subjects used in the previous studies. Some of the subjects assigned to the reading comprehension group had participated in the earlier study using the Modified Rhyme Test. However, none of the subjects in the listening comprehension group had been in any of the prior intelligibility or word recognition tests using synthetic speech.

Stimuli. Fifteen narrative passages were chosen more-or-less randomly from several published adult reading comprehension tests. The exact details of the passages and their original sources are provided in Table I below. An example of one of the passages is provided in Appendix D.

Each passage was initially typed in orthographic form with punctuation into a text file. These files were then used as input to the text-to-speech system and as a source for preparing the typed versions of the passages used in the reading comprehension condition. All fifteen passages were recorded on audio tape at a speaking rate in excess of 190 words/minutes for later playback. Two sets of response booklets were prepared, one for the listening group and one for the reading group. The booklets, which contained a varying number of multiple-choice questions keyed to each paragraph, were arranged in order according to the presentation schedule of the paragraphs on the audio tape. The booklets for subjects in the reading group also

TABLE I
Characteristics of the Passages used
to Measure Comprehension



Passage	No. of words	Duration (secs)	No. of Test Questions	General topic	Source*
01	212	75	6	lens buying	Coop English
02	159	56	4	measuring distance to nearby stars	Coop English
03	327	135	8	language	Iowa
04	198	75	4	retail institutions	Nelson-Denny
05	175	70	4	noise pollution	Nelson-Denny
06	204	82	4	geology	Nelson-Denny
07	206	68	4	philosophy	Nelson-Denny
08	207	80	4	radioactive dating	Nelson-Denny
09	292	117	8	history	Iowa
10	315	100	9	sea	Iowa
11	265	101	7	New Mexico	Stanford
12	322	125	6	Fox hunting	Stanford
13	253	98	6	Claude Debussy	Stanford
14	267	107	7	Aluminum	Stanford
15	212	82	6	Roger Bannister	Stanford

* The full references to these reading tests can be found in the reference section at the end of the chapter.

included a typed copy of the passage immediately before the appropriate set of questions. Appendix E provides the set of questions corresponding to the passage given in Appendix D.

Procedure. Half of the forty-four subjects were assigned to the listening group and the other half to the reading group.

Subjects assigned to the reading group were tested together in a classroom while the subjects in the listening group were tested in small groups of five or six subjects each using the listening facilities of the previous studies. These subjects wore headphones and listened to the passages under the same conditions as the earlier subjects.

Instructions to the subjects in both groups emphasized that the purpose of the study was to evaluate how well individuals could comprehend and understand continuous synthetic speech produced by a reading machine. Subjects in the listening group were told that they would hear narrative passages about a wide variety of topics and that their task was to answer the multiple-choice questions that were keyed to the particular passages as best as they could based on the information contained in the passages they heard. Similar instructions were provided to the reading comprehension group.

As in the previous word recognition study, the listening comprehension group was presented with test passages in a self-paced format with the experimenter present in the testing room operating the tape recorder via remote control. A given test passage was presented only once for listening, after which subjects immediately turned their booklets to the appropriate set of test questions.

The subjects in the reading comprehension group were permitted to read each passage only once and were explicitly told that they should not go back over the passage after reading it or while answering the questions. This procedure was a departure from the typical methods used in administering standarized reading comprehension tests. Usually, the test passage is available to the subject for inspection and re-reading during the entire testing session. However, for present purposes, we felt that comparisons between reading and listening comprehension might be more closely matched by limiting exposure to one pass through the materials.

The subjects in both groups were told at the beginning of testing that the first two passages of the test and the accompanying questions were only for practice to familiarize them with the materials and nature of the test format. These two passages were not scored in the final analyses reported here.

Results and Discussion

The multiple-choice questions for each of the thirteen test passages were scored separately for each subject. A composite score was then obtained by simply cumulating the individual scores for each passage and then expressing this value as a percentage of the total possible score across all of the passages.

The overall results for both reading and listening comprehension are shown in Figure 3 summed over all thirteen test passages. The data are also broken down in this figure by 1st and 2nd half of the test.

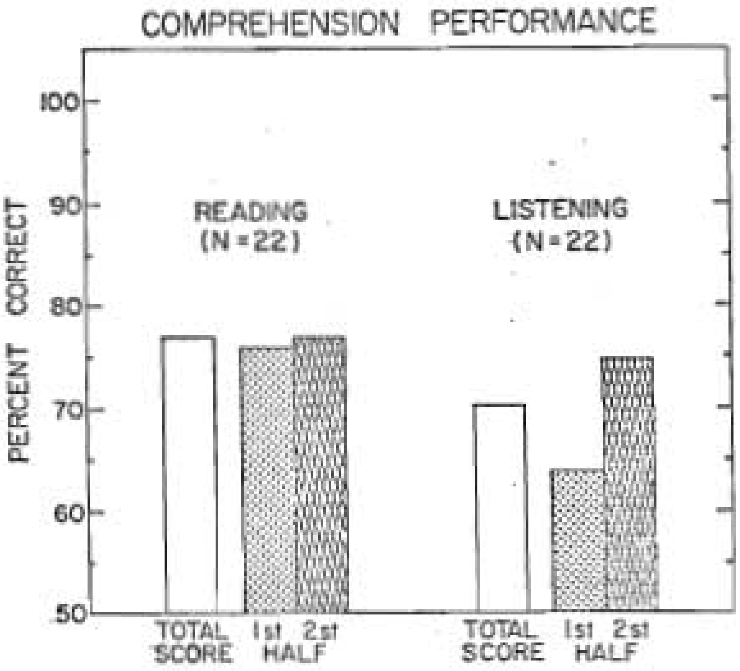


Figure 3. Percent correct comprehension scores for reading and listening groups.

The average percent correct was 77.2% for subjects in the reading comprehension group and 70.3% for subjects in the listening comprehension group. The 7% difference between these two means is small but statistically significant by a t-test for independent groups (p < .05).

Although the reading comprehension group showed better performance overall when compared with the listening comprehension group, a breakdown of the comprehension scores for the two halves of the test showed a significant (p < .001) improvement in performance only for the subjects in the listening comprehension condition. There were no differences between first and second halves of the test for subjects in the reading comprehension group. The finding of improved performance in the second half of the test for subjects in the listening group is consistent with our earlier observations in the word recognition tests showing that listening performance improves for synthetic speech after only a short period of exposure. When the two comprehension groups are compared on the same passages in the last half of the test, their performance is equivalent ($p > .05$) which suggests that the overall difference between the two groups is probably due to familiarity with the output of the synthesizer and not due to any inherent difference in the basic strategies used in comprehending or understanding the content of these passages. This conclusion is strengthened even further by the fact that the thirteen passages are correlated across both testing conditions. In this case, a very high correlation ($r= +.97$) was observed between reading and listening comprehension scores for individual passages. Passages that are difficult to comprehend when read are also difficult to comprehend when listened to and vice versa. The time taken to complete all passages in both tests was, however, roughly the same, lasting between 45 and 50 minutes.

After the listening comprehension test was completed, we solicited additional subjective evaluations of the speech produced by the synthesizer and the nature of the comprehension test itself. Twenty of the twenty-two subjects indicated that they were able to comprehend and understand the content of the passages "well" or "very well." Only two of the subjects reported difficulty in comprehension and even these two did not indicate that they were merely guessing, an available response alternative.

Several of the subjects reported improved ability to understand the speech as testing progressed. Others described several problems in the quality of synthesis, the location of pauses, the existence of inappropriately stressed words and, the occasional presence of very long "run-on" sentences in several passages. Finally, several other subjects suggested that each test passage should be presented twice so they could review some of the specific details and facts that were stated explicitly. For the most part, however, the subjects found listening to the speech interesting and felt that they had performed reasonably well in comprehending the passages. None of the subjects reported any major distractions in the quality of the synthetic speech that interfered with their ability to attend to or understand the content of the passages. Thus, subjects are able to adapt easily to relatively long passages of synthetic speech with little exposure or practice.

Conclusions

The results of the comprehension test indicate that naive subjects are able to comprehend synthetically produced spoken passages of narrative text output from an unrestricted text-to-speech system. Their performance is roughly comparable to subjects who have been asked to read the same passages of text and answer the same questions. As in the case of our other tests using synthetic speech, there appears to be an initial period during which subjects are simply becoming familiar with the quality of the synthesizer, the prosodic rules of the system and the style of the material. Even after only a few minutes of exposure, comprehension performance improves substantially and eventually approximates levels observed when subjects read the same passages of text.

It should also be pointed out here that the comprehension performance observed in these tests was obtained with a reading rate in excess of 180 words per minute. This rate is about the rate at which people typically speak in normal conversations or when they read text aloud. The present results therefore suggest that it is not necessary to slow down the speaking rate or adjust the synthesis to obtain relatively high levels of listening comprehension for continuous text. Until the present tests were carried out, it was assumed by some investigators that synthetic speech had to be output at a much slower rate to maintain intelligibility and therefore facilitate comprehension.

Based on the results of the present comprehension test as well as the other tests of intelligibility and word recognition that were carried out, there is good reason to believe that the basic design of the MIT text-to-speech system is valid. The system can produce not only highly intelligible synthetic speech, as shown in our earlier tests, but the quality of the synthetic speech can be understood and comprehended at reasonably high levels. While there are no doubt many subtle details of the system that might be improved, the results of these preliminary tests support the general conclusion that very high quality synthetic speech can be produced automatically from unrestricted text and that such a system could be implemented in applied settings in the immediate future. After some thirty years of research, the widespread use of text-to-speech and voice response systems in computer aided instruction and as aids for the handicapped is now a realistic goal. The obstacles are no longer questions of research into the basic principles of speech production, perception, and linguistic analysis but are simply the practical matters of implementation and economics.

13.5 GENERAL DISCUSSION AND CONCLUSIONS

The results of the three tests designed to evaluate intelligibility, word recognition, and listening comprehension indicated very high levels of performance for the current version of the text-to-speech system. While these tests are only preliminary, they have provided an initial benchmark against which to compare the performance of the present system with other text-to-speech systems. Moreover, the present results have provided a basis for evaluating the overall design of the system and the functioning of several of the individual components. Since a relatively large amount of text was specifically generated for this project, we were able to identify a number of errors in the operation of the system which ordinarily might not have been detected. In this last section of the chapter, we summarize briefly a few of the errors we were able to uncover during and after the evaluation. We will also point out some of the limitations of the current evaluation results and then discuss several directions for additional testing in the future.

After the test materials for the evaluation project were generated, it was possible to go back and examine the output of each module individually in order to determine whether it provided a correct analysis of the input text. Errors of various kinds in the final spoken output could originate at several different modules in the system. In addition, there also could be errors resulting from transcription that we would not associate with the operation of the text-to-speech system itself.

Of all of the errors observed, we discovered only one that could legitimately be classified as a transcription error. In this case, the word "harmonies" was incorrected typed into the system as "harmonics" and was not detected in subsequent proof reading. All remaining errors could be located at one or more modules of the system. These errors consisted of incorrect parsings, pronunications or stress assignments. An error located at one module often affected analyses carried out by other modules. Sometimes the results of these errors were quite noticeable in the spoken output, particularly when the errors produced segmental distinctions that could be detected in pronunciation. However, in other cases, particularly where stress assignment was involved, the differences were more difficult to detect.

At the time this report was completed, we were able to locate only two errors in the operation of the first module of the system. Recall that this module (FORMAT) has a dictionary that converts abbreviations, symbols and numbers to words for subsequent processing. One error involved the abbreviation U.S. in which a space was incorrectly typed between U. and S. The rule which was applied here places an end-of-sentence period in the output if an abbreviatory period (as in U.) is followed by one or more spaces and a capital letter (the S). Thus two sentences were formed, one ending in "U." and the other beginning with "S." This error causes an incorrect pitch contour to be placed on the output as well as inappropriate segmental durations to be assigned in later modules.

Another error involved the abbreviation 19th. In all cases, alphanumerics are spelled out completely by this module. For example, 19th was pronounced as "one-nine-T-H" on output. In words such as 19th or 100-yard, the alphabetic and numeric sections are separable and could be pronounced. However, in a true alphanumeric such as 103S or a3c, it is correct to spell out all of the symbols.

A number of errors were also detected in the module DECOMP which is responsible for decomposing words into morphs by reference to the morph lexicon. In several cases, the wrong morphs were identified resulting in perceptible segmental errors in the speech output. In other cases, the correct morphs were obtained, but the stress assignment of the constituent morphs was different for the morphs in isolation than for the morphs when concatenated in a polymorphemic word. We also identified several words that should have been in the lexicon since their pronunication could not be handled by the existing spelling-to-sound rules.

Several errors in the operation of the spelling-to-sound rules were also detected. These errors resulted in the wrong pronunciation which was quite noticeable in listening. For example, the second syllable of the word "Britain" was pronounced like the second syllable in the word "maintain."

In a number of other cases, we were able to identify problems in the operation of the parser, particularly in recognizing the correct part of speech. For example, the word "close" can be either an adjective or verb each with a different pronunciation. Several problems were also observed with the word "affect" which can be either a noun or a verb. In each of these cases, the part of speech was incorrectly identified by the parser resulting in the wrong choice in pronunciation on output.

Finally, there were several cases, especially with the Haskins anomalous sentences, in which the parser incorrectly assigned the verb (which could also be a noun) to the previous noun phrase. This error is not surprising since the parser has a basic preference for noun phrases anyway when a choice is available. However, this often produced inappropriate sentence stress resulting from incorrect pitch and segmental durations. In some cases, these differences could be readily observed whereas in others the effects were substantially more difficult to detect even with careful and repeated listening. These observations are consistent with an earlier perceptual study of the durational rules carried out by Carlson, Granstrom and Klatt (1979). They found that a deletion of a phrase boundary produced only negligible effects on listeners' evaluations of the naturalness of synthetic speech.

Some of the errors described above are considered to be relatively minor and can be corrected rather easily by the simple addition of polymorphemic entries in the morph lexicon. Since this evaluation was completed, a "pre-parser" has been implemented which corrects a number of the parsing errors in which the sentential verb was included in the preceding noun phrase. However, some of the other parsing errors are not as easy to correct. Errors made by the first module and the spelling-to-sound rules are highly context-dependent and are not easily amenable to simple change by rule. From our examination of the errors uncovered so far, all cases could be accounted for and located in some module for the system. There were no errors detected which escaped explanation at the present time although further study is continuing.

The results of the present evaluation study have several limitations and these should be summarized here briefly for future reference. First, we did not carry out any of the control conditions for the three types of tests using natural speech. To some extent this might be considered an important addition and extension of the current evaluation since it is the level of performance with natural speech that is frequently used as the yardstick against which to compare the quality of synthetic speech. There can be little doubt that tests with natural speech would show higher levels of performance when compared with synthetic speech. But it should be emphasized here that the levels of performance in the current study are already quite high to begin with and therefore it is not immediately obvious what would be gained from such additional tests with natural speech.

Secondly, with regard to measuring intelligibility of the segmental output, it is clear that the Modified Rhyme Test is much too easy for listeners, even naive listeners, and additional tests using an open response set should be employed. Additional testing under varying noise conditions may also provide further information concerning the quality of the synthesis and its resistance to noise and distortion. In this regard, the analysis of the Haskins anomalous sentences should also provide us with a rich source of data on phonetic confusions using an open response set. We are planning additional detailed analyses of these data.

Finally, the comprehenssion test that we used was relatively gross in its ability to distinguish between new knowledge acquired from listening to text and knowledge obtained from inferences drawn at the time of comprehension or later at the time of testing. Of course, this is a problem related more to several broader issues in language comprehension and understanding than to questions surrounding text-to-speech and speech synthesis by rule. Nevertheless, it may be possible to learn a great deal more about language comprehension and the interaction between top-down and bottom-up knowledge sources in speech perception by the advances that have been made in conceptualizing various linguistic problems within the context of a functional text-to-speech system. The success of the current system and its capabilities to process unrestricted text must be traced at least, in part, to the existence of an explicit model of the underlying linguistic structure that is common to both text and speech and to the rule systems relating the two domains.

In summary, the results of our evaluation tests designed to measure phoneme intelligibility, word recognition and comprehension of synthetic speech produced by the MIT text-to-speech system have demonstrated good to excellent performance on a wide range of materials. No major problems were uncovered in the design of the system nor were any serious errors identified in any of the component modules of the system to date. The present results, although preliminary, support the general conclusion that very high quality synthetic speech can be produced automatically from unrestricted English text and that such a system could be implemented in an applied setting in the very near future.

Acknowledgements

The research described in this chapter was carried out while the author was a Guggenheim Fellow at the Research Laboratory of Electronics, MIT. The work was supported, in part, by NIMH grant MH-24027-05 and NIH grant NS-12179-04 to Indiana University in Bloomington. Special thanks are due Sharon Hunnicutt for her efforts in preparing the materials used in these experiments and her interest and advice in all phases of the project. Her expertise and intimate knowledge of the design and operation of the text-to-speech system was invaluable to me during the preparation of this report. Additional assistance was provided by Tom Carrell, Wendy Crawford and Amanda Walley who carried out the initial data analyses. Robert E. Remez, Dennis Klatt and Jon Allen also provided helpful comments at various stages of this project and their assistance is appreciated.

References

Allen, J. Reading Machines for the Blind: The Technical Problems and the Methods Adopted for Their Solution. IEEE Transactions on Audio and Electroacoustics, 1973, Vol. AU-21, No. 3, 259-264.

Allen, J. Synthesis of Speech from Unrestricted Text. Proceedings of the IEEE, 1976, 64, 4, 433-442.

Allen, J. Hunnicutt, S., Carlson, R. and Granstrom, B. MITalk-79: The 1979 MIT Text-to-Speech System. In J. J. Wolf and D. H. Klatt (Eds.) Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America, New York: Acoustical Society of America, 1979, Pp. 507-510.

Carlson, R., Granstrom, B. and Klatt, D.H. Some Notes on the Perception of Temporal Patterns in Speech. Proceedings of the Ninth International Congress of Phonetic Sciences. Volume II. Copenhagen: Institute of Phonetics, University of Copenhagen, 1979, Pp. 260-267.

Carlson, R., Granstrom, B. and Larsson, K. Evaluation of a Text-to-Speech System as a Reading Machine for the Blind. Speech Transmission Laboratory, QPSR 2-3, (1976) Pp. 9-13.

Cooperative English Tests: Reading Comprehension. Form 1B. Princeton, N.J.: Educational Testing Service, 1960.

Egan, J. P. Articulation Testing Methods. Laryngoscope, 1948, 58, 955-991.

Fairbanks, G. Test of Phonemic Differentiation: The Rhyme Test. Journal of the Acoustical Society of America, 1958, 30, 596-600.

Iowa Silent Reading Tests. Level 3. Form E. New York: Harcourt Brace Jovanovich, 1972.

House, A. S., Williams, C. E., Hecker, M. H. L. and Kryter, K. D. Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set. Journal of the Acoustical Society of America, 1965, 37, 158-166.

Ingeman, F. Speech Synthesis by Rule Using the FOVE Program. Haskins Laboratories Status Report on Speech Research, SR-54, (1978), Pp. 165-173.

Marslen-Wilson, W. D. and Welsh, A. Processing Interactions and Lexical Access During Word Recognition in Continuous Speech. Cognitive Psychology, 1978, 10, 29-63.

Miller, G. A., Heise, G. and Lichten, W. The Intelligibility of Speech as a Function of the Context of the Test Materials. Journal of Experimental Psychology, 1951, 41, 329-335.

Miller, G. A. and Isard, S. Some Perceptual Consequences of Linguistic Rules. Journal of Verbal Learning and Verbal Behavior, 1963, 2, 217-228.

The Nelson-Denny Reading Test. Form D. Boston: Houghton-Mifflin, 1973.

Nye, P.. W. and Gaitenby, J. Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results). Haskins Laboratories Status Report on Speech Research, SR-33, (1973), Pp. 77-91.

Nye. P. W. and Gaitenby, J. The Intelligibility of Synthetic Monosyllable Words in Short, Syntactically Normal Sentences. Haskins Laboratories Status Report on Speech Research, SR-37/38, (1974), Pp. 169-190.

Nye, P. W., Ingeman, F. and Donald, L. Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences Among Different Speech Forms. Haskins Laboratories Status Report on Speech Research, SR-41, (1975), Pp. 117-126.

Pisoni, D. B. Speech Perception. In W. K. Estes (Ed.) Handbook of Learning and Cognitive Processes Volume 6. Hillsdale, N. J. Lawrence Erlbaum Associates, 1978, Pp. 167-233.

Stanford Test of Academic Skills: Reading. College Level II-A. New York: Harcourt Brace Jovanovich, 1972.

APPENDIX 4

Sample Test Trials from the Modified Rhyme Test

1.	a) bad	b) back	c) ban	d) bass	e) bat	f) bath
2.	a) beam	b) bead	c) beach	d) beat	e) beak	f) bean
3.	a) bus	b) but	c) bug	d) buff	e) bun	f) buck
4.	a) case	b) cave	c) cape	d) cane	e) cake	f) came
5.	a) cuff	b) cut	c) cuss	d) cub	e) cup	f) cud
6.	a) dip	b) din	c) dill	d) dig	e) dim	f) did
7.	a) dub	b) dun	c) dung	d) dug	e) duck	f) dud
8.	a) fizz	b) fin	c) fill	d) fig	e) fib	f) fit
9.	a) hear	b) heath	c) heal	d) heave	e) heat	f) heap
10.	a) kid	b) kit	c) kill	d) kin	e) king	f) kick
11.	a) lace	b) lame	c) lane	d) lay	e) lake	f) late
12.	a) man	b) math	c) mad	d) mat	e) mass	f) map
13.	a) pace	b) pane	c) pave	d) page	e) pay	f) pale
14.	a) path	b) pat	c) pack	d) pad	e) pass	f) pan
15.	a) peas	b) peak	c) peal	d) peace	e) peach	f) peat
16.	a) pip	b) pick	c) pin	d) pill	e) pit	f) pig
17.	a) puff	b) pus	c) pub	d) pun	e) puck	f) pup
18.	a) rate	b) race	c) ray	d) raze	e) rave	f) rake
19.	a) safe	b) sake	c) same	d) sane	e) save	f) sale
20.	a) sat	b) sag	c) sack	d) sap	e) sass	f) sad
21.	a) seed	b) seek	c) seen	d) seep	e) seem	f) seethe
22.	a) sill	b) sick	c) sing	d) sit	e) sin	f) sip
23.	a) sup	b) sud	c) sun	d) sum	e) sub	f) sung
24.	a) tap	b) tang	c) tam	d) tan	e) tab	f) tack
25.	a) tease	b) tear	c) teak	d) teal	e) team	f) teach

APPENDIX B

Sample Test Materials from the Harvard Psychoacoustic Sentences

1. The birch canoe slid on the smooth planks
2. Glue the sheet to the dark blue background
3. It's easy to tell the depth of a well
4. These days a chicken leg is a rare dish
5. Rice is often served in round bowls
6. The juice of lemons makes fine punch
7. The box was thrown beside the parked truck
8. The hogs were fed chopped corn and garbage
9. Four hours of steady work faced us
10. A large size in stockings is hard to sell
11. The boy was there when the sun rose
12. A rod is used to catch pink salmon
13. The source of the huge river is the clear spring
14. Kick the ball straight and follow through
15. Help the woman get back to her feet
16. A pot of tea helps to pass the evening
17. Smoky fires lack flame and heat
18. The soft cushion broke the man's fall
19. The salt breeze came across from the sea
20. The girl at the booth sold fifty bonds
21. The small pup gnawed a hole in the sock
22. The fish twisted and turned on the bent hook
23. Press the pants and sew a button on the vest
24. The swan dive was far short of perfect
25. The beauty of the view stunned the young boy

APPENDIX C

Sample Test Materials from the Haskins Anomalous Sentences

1. The wrong shot led the farm
2. The black top ran the spring
3. The great car met the milk
4. The old corn cost the blood
5. The short arm sent the cow
6. The low walk read the hat
7. The rich paint said the land
8. The big bank felt the bag
9. The sick seat grew the chain
10. The salt dog caused the show
11. The last fire tried the nose
12. The young voice saw the rose
13. The gold rain led the wing
14. The chance sun laid the year
15. The white bow had the bed
16. The near stone thought the ear
17. The end home held the press
18. The deep head cut the cent
19. The next wind sold the room
20. The full leg shut the shore
21. The safe meat caught the shade
22. The fine lip tired the earth
23. The plain can lost the men
24. The dead hand armed the bird
25. The fast point laid the word

APPENDIX D

Sample Passage Used to Test Listening Comprehension

The lens buyer must approach the problem of purchasing a lens of large aperture with caution. The first question to consider is whether the work one intends doing will actually require the extreme speed afforded by such a lens. Despite the glowing advertising claims, no extremely rapid lens is capable of giving, even when stopped down to its best aperture, the sharpness of definition which may be obtained with a well-corrected (and much lower-rpiced) lens of smaller maximum aperture. It is very doubtful if there exists a lens with maximum aperture in excess of F 4.5 which will give really sharp definition, whether wide open or at any smaller opening; the deficiencies of the large-apertured lens, if it is a fairly good one, will not be noticed in small contact prints; but in pictures enlarged to any considerable extent, they will be evident (or examination of the print with the low magnification of a reading glass will make them evident). With modern, extremely rapid films and with synchronized flash available for the amateur who can afford to go in for the type of photography that requires this kind of equipment, the occasions are indeed rare when a lens faster than F 4.5 is really needed.

APPENDIX E

Test Questions for the Comprehension Passage Shown in Appendix D

1. The main thought of this passage is that
   A good photographic work requires the use of a fast lens
   B lenses of small aperture provide less sharpness than lenses of large aperture
   C lenses of small aperture are to be preferred for most photographic work
   D modern photographic equipment requires the use of lenses of large aperture

2. We may infer that some advertisements for photographic lenses tend to recommend the purchase of
   A large-aperture lenses
   B small-aperture lenses
   C the most appropriate lenses
   D F 4.5 lenses

3. As the aperture of a lens is increased, the
   A price tends to decrease
   B speed of the lens tends to decrease
   C sharpness of its focus tends to decrease
   D speed of the lens tends to remain constant

4. The writer's attitude toward the advertising materials which are mentioned is one of
   A indifference B disbelief
   C acceptance D enthusiasm

5. The writer's main purpose is to
   A encourage the purchase of fast lenses
   B discourage the use of synchronized flash
   C encourage the use of rapid films
   D discourage the purchase of fast lenses

6. To obtain pictures of maximum sharpness, the writer strongly recommends the use of
   A lenses of large aperture
   B lenses of small aperture
   C lower priced films
   D contact prints

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 5 (1978-79) Indiana University]

Discrimination of relative onset time of two-component tones by infants[1]

Peter W. Jusczyk

Dalhousie University, Halifax, Nova Scotia

David B. Pisoni

Indiana University, Bloomington, Indiana

Amanda Walley and Janice Murray

Dalhousie University, Halifax, Nova Scotia

Short Title: Discrimination of relative onset-time by infants

Abstract

A great deal of research has focused on the perception of voice onset time (VOT) differences in stop consonants. Yet, the nature of the mechanisms responsible for the perception of these differences is still the subject of much debate. Recently Pisoni (1977) has presented evidence which suggested that the perception of VOT differences by adult listeners may reflect a basic limitation on processing temporal order information by the auditory system. For adults, stimuli with onset differences approximately greater than 20 msec. are perceived as successive events (either leading or lagging), while stimuli with onset differences less than about 20 msec. are perceived as simultaneous events. Thus, differences in voicing may have an underlying perceptual basis in terms of three well-defined temporal attributes corresponding to leading, lagging or simultaneous events at onset. The present experiment was carried out to determine whether young infants can discriminate differences in temporal order information in nonspeech signals and whether their discrimination performance parallels the earlier data obtained with adults. Discrimination was measured with the high amplitude sucking (HAS) procedure. The results indicated that infants can discriminate differences in the relative onset of two events; the pattern of discrimination also suggested the presence of three perceptual categories along this temporal continuum although the precise alignment of these categories differed somewhat from the values found in the earlier study with adults.

I. INTRODUCTION

A considerable body of psychophysical evidence has accumulated in the last few years on the perception of nonspeech signals that have properties similar to those found in speech (Cutting & Rosner, 1974; Cutting, Rosner & Foard, 1976; Miller, Pastore, Wier, Kelley & Dooling, 1976; Pisoni, 1977). The results of these experiments have shown that the mechanisms used in speech perception appear to be constrained in numerous principled ways by the basic capabilities of the auditory system to process incoming sensory information (Searle, Jacobson & Rayment, 1979; Miller, Engebretson, Spenner & Cox, 1977). These findings mesh nicely with the theoretical work of Stevens (1972) who suggests that the constraints imposed on acoustic signals by the auditory system may initially delineate some of the basic kinds of acoustic events and properties that languages have exploited in realizing phonetic distinctions.

During this period of time there has also been a great deal of research dealing with the perception of voice onset time (VOT) in stop consonants by human adults and infants as well as animals such as chinchillas (Kuhl & Miller, 1975; 1978), and monkeys (Morse & Snowdon, 1975; Sinnott, Beecher, Moody & Stebbins, 1976; Waters & Wilson, 1976). However, despite the prevalent interest in the perception of VOT, the precise nature of the sensory and perceptual mechanisms responsible for these findings from seemingly diverse organisms is still a matter of some controversy among numerous investigators (Stevens & Klatt, 1974; Lisker, 1975; Miller _et al._, 1976; Summerfield & Haggard, 1977; Kuhl & Miller, 1978).

represented by variations in voice onset time (VOT). Indeed, Stevens and Klatt (1974) have even remarked that the inventory of phonetic features found in natural languages seems to consist of the presence or absence of sets of acoustic attributes or cues rather than simply continuous changes in a small set of parameters or dimensions, a view that is similar in spirit to the founders of distinctive feature theory (Jacobson, Fant & Halle, 1952). One such set of distinctive attributes that may be coded by the auditory system could be temporal information about the timing of laryngeal and supralaryngeal events in the production of stop consonants.

To study the underlying perceptual basis of the voicing feature, Pisoni (1977) used a set of nonspeech stimuli differing in the relative onsets of two component tones of different frequencies, a temporal dimension known to be an important acoustic cue to the perception of voicing in stops. Earlier experiments with synthetic speech stimuli had established the importance of the so-called F1 "cutback" cue as a perceptual dimension to voicing in stops so there was sufficient justification for focusing on the same temporal variable in these nonspeech stimuli (Liberman, Delattre & Cooper, 1958). The results obtained in identification and discrimination experiments with these tone-onset-time (TOT) stimuli were quite similar to the results observed earlier with synthetic speech stimuli differing in VOT (Lisker & Abramson, 1967; Abramson & Lisker, 1967). Subjects were able to consistently identify these nonspeech stimuli into well-defined perceptual categories. Moreover, discrimination of pairs of these stimuli was very nearly categorical, with performance close to chance for pairs of stimuli selected from within a perceptual category and excellent for pairs of stimuli selected from different perceptual categories. Furthermore, in other experiments involving categorization and temporal order judgments, it was possible to identify a basis for the underlying perceptual categories found with these nonspeech stimuli in terms of whether the acoustic events at stimulus onset were perceived as simultaneous or successive, and if the latter, whether the temporal order of the component events could be identified as leading or lagging. These three properties of stimulus onsets -- lead, lag and simultaneity have been found to characterize the major differences in voicing among stops in a large number of languages as represented by the VOT dimension (Lisker & Abramson, 1964, 1967). Thus, it seemed likely that these perceptual results with nonspeech stimuli could be used as an account of the perceptual findings obtained with speech stimuli differing in VOT.

The temporal order hypothesis of voicing perception proposed by Pisoni (1977) at that time was also able to account for a seemingly diverse set of findings on the perception of VOT that had been reported in the literature over the last few years. For example, it had been known for some time that cross-language differences exist in the perception of VOT by adults (Lisker & Abramson, 1967). Moreover, a number of perceptual experiments were also carried out on the discrimination of VOT by infants, chinchillas and monkeys indicating a strong possibility of a psychophysical or sensory basis for the observed discrimination data. These somewhat diverse results could be accommodated by simply postulating a common underlying basis for the discrimination involving a basic constraint on the auditory system's ability to respond to differences in temporal order between two events at onset.

Recent interest in the basic sensory capabilities of the auditory system has also provided additional information about the underlying psychophysical basis of categorical perception, a finding once thought to be unique only to the perception of speech sounds. Several studies have demonstrated that categorical perception is not confined exclusively to the perception of speech signals per se, but instead may be a very general characteristic of the way sensory systems respond to changes in one component of a complex stimulus when other properties of the stimulus remain constant (Pastore, 1976; Pastore, Ahroon, Buffuto, Friedman, Puleo & Fink, 1977). Moreover, the prevalent view that categorical perception of speech was primarily a consequence of identification or labeling brought about through phonetic categorication has now been seriously questioned by the demonstration of marked changes in sensitivity (d') and bias in the region corresponding to the boundary separating perceptual categories (Wood, 1976). Thus, these results imply that the perceptual categories employed in the phonological systems of languages may have a natural and well-defined basis in terms of what is known about the sensory capacities of the auditory system itself, above and beyond considerations related to the interpretation of these acoustic signals of speech.

Taken together, these recent studies promote the general view that many of the basic functions and mechanisms of the auditory system are used in processing both speech and nonspeech signals alike. While there is no doubt important differences in perception between speech and nonspeech signals, there may also be many similarities based on common psychophysical processes that could help to specify the exact sensory and perceptual basis of the acoustic correlates of distinctive features that occur in speech. Such perceptual considerations may also be relevant to explanations of numerous phonetic and phonological processes that seem to occur universally in language (Lieberman, 1976).

In addition to the theoretical interest in the possible sensory and perceptual correlates of distinctive features in speech, the recent findings on the perception of nonspeech signals differing in relative onset time are also relevant to several well-known findings in perceptual development, particularly the demonstration by Eimas _et al_. (1971) that one-month old infants perceive differences in VOT categorically. These results as well as a number of other findings with infants have been interpreted as evidence for the existence of a "speech mode" of perception and the operation of "specialized" perceptual mechanisms for processing speech signals in humans (for a review see Eimas, 1978; Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967).

In the initial study involving stimuli differing in VOT, Eimas _et al_. (1971) demonstrated that infants could discriminate between two speech sounds selected from across an adult phoneme boundary but could not discriminate two stimuli selected from within the same adult perceptual category even though the acoustic differences between the stimuli were equal, at least in terms of the physical dimension of VOT. These results were quite provocative at the time suggesting that infants might have access to mechanisms of phonetic categorization at an extremely early age. Moreover, these results were interpreted by Eimas and others as support for the idea that the mechanisms responsible for categorical perception of speech sounds might be specified innately in humans.

One of the most important claims of these early infant experiments on VOT was the asertion that the infants were responding to these speech signals in a "linguistically relevant manner" that involved the phonetic coding of these stimuli into abstract perceptual categories comparable to those observed in adult subjects. An alternative view -- that these infants were simply responding to the psychophysical differences between these signals in the absence of explicit phonetic categorization, was proposed by Stevens and Klatt (1974) in light of the results they obtained in several perceptual experiments with adults. These investigators argued that the infants in the Eimas _et al_. experiments were simply responding to the presence or absence of a voiced F1 formant transition at onset rather than to VOT per se. In a reply to this paper, Lisker (1975) has shown that it is primarily F1 onset frequency that adult listeners respond to as a positive cue voicing rather than the F1 frequency shift observed by Stevens and Klatt. Summerfield and Haggard (1977) have confirmed and extended Lisker's earlier findings in a series of experiments that systematically varied both spectral and temporal cues to voicing. Although these perceptual experiments have provided useful information about the numerous cues to voicing in stops and their potential interactions, the data were all collected with adult subjects who no doubt had a very long history in mastering English phonology. Thus, the claim that infants are responding to VOT differences in a linguistically relevant manner still remains largely unresolved.

The results of two cross-language experiments using the same stimuli differing in VOT have also provided additional evidence that young infants can discriminate differences in this acoustic dimension.

Moreover, the results have been interpreted as support for the claim that infants are sensitive to three primary modes of voicing in stop consonants. In one study, Lasky, Syrdal-Lasky and Klein (1975) studied four to 6½ month-old infants born to Spanish-speaking parents and found evidence suggesting the presence of three voicing categories in their discrimination data. One area of high sensitivity occurred in the region of +20 to +60 msec. which corresponds to the English voiced-voiceless distinction, whereas the other area of high sensitivity occurred in the region between roughly -20 and -60 msec. These discrimination results are interesting because Spanish has only one phoneme boundary separating its voiced and voiceless stops and that boundary does not coincide with either of the two boundaries that Lasky *et al*. inferred from their discrimination data. The apparent discrepancy between the adult and infant data suggests that the infants in this study were probably responding to some set of acoustic attributes or cues in these VOT stimuli independently of their phonetic status or exposure to them in the language learning environment.

In another study Streeter (1976) found that Kikuyu infants also showed evidence of discriminating three categories of voicing for labial stops. Her results are also of some importance, because there are no voicing contrasts for labial stops in Kikuyu, although there are voicing contrasts for stops at other places of articulation in this language. Since this particular contrast was not phonologically distinctive in the adult language, and therefore probably occurred quite infrequently in the language learning environment of these infants, the infants' discrimination of VOT must have been entirely based on the acoustic and psychophysical attributes of the stimuli themselves. This conclusion is strengthened by the fact that the regions of high discriminability found in this study were similar to those obtained in the earlier study by Lasky _et al_. despite the differences between the two languages.

The results of both cross-language investigations of the perception of voicing in young infants as well as the initial findings of Eimas _et al_. indicate that young infants can discriminate differences in VOT. However, the underlying basis of the infants' discrimination performance may simply be a consequence of the presence of psychophysically defined regions of high discriminability that exist in the VOT continuum itself rather than processes that involve phonetic categorization or interpretation of these signals as speech. A clear precedent of this notion already exists in an earlier study of infants' perception of nonspeech stimuli conducted by Jusczyk, Rosner, Cutting, Foard and Smith (1977). These investigators found evidence that infants' discrimination of sinewave stimuli differing in rise-times was categorical. This demonstration that infants display categorical discrimination of nonspeech as well as speech sounds, supports the view that the infants' perceptual behavior in these situations may be the consequence of mechanisms attuned to psychophysical properties in the acoustic signal. Thus, the infants in the previous VOT studies may not have perceived these signals linguistically as Eimas _et al_. have claimed, but instead may have been responding to some complex set of psychophysical properties that separates each of the three primary modes of voicing. One such property of these VOT stimuli may be the relative timing of the component events at stimulus onset.

If the auditory system responds to temporal order information in both speech and nonspeech signals in terms of coding simultaneous and successive events as salient perceptual attributes, we would expect to find that such mechanisms are also present and operative in young infants given the earlier results on the discrimination of VOT summarized above. Moreover, such an outcome in young infants would be consistent with the nonspeech results of Jusczyk _et al_. (1977) and with predictions based on the nonspeech results obtained with adults by Pisoni (1977). The present experiment was therefore carried out to determine whether infants can discriminate differences in temporal order in nonspeech signals having properties similar to those found in speech. In addition, we were also interested in determining whether the pattern of discrimination along this nonspeech continuum would be comparable to that found earlier in adult subjects.

II. METHOD

1. _Procedure_

Each infant was tested individually in a small laboratory room. The infant was placed in a reclining chair which faced a rear projection screen approximately .5 m away. An image of a man was displayed on the screen for the entire test session. The projection screen was situated just above a loudspeaker through which the test stimuli were played. Each infant sucked on a blind nipple held in place by an experimenter who wore headphones and listened to recorded music throughout the test session. A second experimenter in an adjacent room monitored the apparatus.

The experimental procedure was a modification of the high-amplitude sucking technique devised by Siqueland and DeLucia (1969). For each infant, the high-amplitude sucking criterion and the baseline rate of high-amplitude sucking were established prior to the presentation of any test stimuli. The criterion for high-amplitude sucking was adjusted so as to produce rates of 15 to 35 sucks/min. After a baseline rate was established, the presentation of stimuli was made contingent upon the rate of high-amplitude sucking. Since the stimuli had a maximum duration of 300 msec. and a 750 msec. interstimulus interval was used, the maximum stimulus presentation rate was approximately one stimulus per second. If the infant produced a burst of sucking responses with interresponse times of less than one second, then each response did not produce one presentation of the stimulus. Rather, the timing apparatus was reset so as to provide continuous auditory feedback for one second after the last response of the sucking burst. Use of a programmable logic board ensured that all stimulus presentations were uninterrupted.

The criterion for satiation to the first stimulus was a decrement in sucking rate of 25% or more over 2 consecutive minutes compared to the rate in the immediately preceding minute. At this point the auditory stimulation was changed without interruption by switching channels on the tape recorder. For infants in the experimental conditions, the change resulted in the presentation of a second acoustically different stimulus. For infants in the control condition, the channels on the tape recorder were switched, but no acoustic change occurred since the same signal had been recorded on both channels of the tape. The postshift period lasted for 4 min. The infant's sensitivity to the change in auditory stimulation was inferred from comparisons of response rates of subjects in the experimental and control conditions during the postshift period.

2. Stimuli

The stimuli were two-tone sequences that were generated digitally on a PDP 11/10 computer with a program that permits the specification of the amplitude and frequency of two sinusoids at successive moments in time (Kewley-Port, 1976). These stimuli were similar to ones used in the earlier experiment by Pisoni (1977). A schematic display of the stimuli is shown in Figure 1. Each stimulus consisted of two tones, a lower one set at 500 Hz and a higher one set a 1500 Hz. The amplitude of the latter was 12 dB lower than the former so that the amplitude relations between the two might parallel those found in a neutral vowel. Both tones were terminated together at the same time. In addition, the duration of the 1500 Hz tone was always held constant at 250 msec. To form the test signals, the duration of the 500 Hz tone was varied systematically in 10 msec. steps from 300 msec. to 160 msec. across the series of stimuli. Thus, the stimuli could be arranged along a temporal continuum according to the degree by which the onset time of 500 Hz tone either led or lagged behind that of the 1500 Hz tone. The endpoint values of this tone-onset-time (TOT) continuum were -70 msec. (in which case the 500 Hz tone leads the 1500 Hz tone by 70 msec.) and +70 msec. (in which case the 500 Hz tone lags behind the 1500 Hz tone by 70 msec.). Digitized waveforms of the stimuli were converted into analog form via a D-A converter, low-pass filtered and then output to a Crown (Model 822) tape recorder in order to prepare the two-channel audiotapes employed in

Insert Figure 1 about here

this experiment.

3. Design

Each infant was seen for one experimental session. Sixteen infants were assigned randomly to each of six test groups. One of these groups (Group I) served as a control condition in which subjects were randomly assigned to one of the 11 two-tone stimuli for the entire session (e.g. +70 vs +70). Subjects in the remaining five test groups were presented with pairs of stimuli differing in tone-onset-time values by 30 msec. The stimulus pairs were chosen so as to permit comparisons of the discriminability of both between-category and within-category contrasts in tone-onset-time. The stimulus values selected for each experimental group are displayed in Table 1. Based on the results of Pisoni's (1977) earlier experiment with adults, Groups II, IV, and VI were presented "Within Category" contrasts of TOT stimuli. For Group II both stimuli were chosen from the "Lead Category". For Group IV, all stimuli were selected from the "Simultaneous Category". For Group VI, stimuli from the "Lag Category" were employed. In contrast, subjects in Groups III and V received stimulus pairings selected from different TOT categories. In the case of Group III, one member of each stimulus pair was selected from the "Lead Category" (i.e. -40 or -30 msec.), and the other from the "Simultaneous Category" (i.e. -10 or 0 msec.). For Groups V, the pairings were between the "Simultaneous Category" (i.e. 0 or +10 msec.) and the "Lag Category" (i.e. +30 or +40 msec.). The presentation order of stimuli was always counterbalanced across subjects for each of the groups.

Insert Table 1 about here

On the basis of these stimulus pairings selected from the earlier adult data, we expected that infants would discriminate only "between category" contrasts that were selected from opposite sides of either the -20 msec. boundary (i.e. lead vs stimultaneous) or the +20 msec. boundary (i.e. simultaneous vs lag). In contrast, we also expected that infants would not discriminate any of the "within category" contrasts that were selected from the same adult perceptual category.

4. Apparatus

A blind nipple was connected to a Grass PTS volumetric pressure transducer which, in turn, was coupled to a Type DMP-4A Physiograph. A Schmitt trigger provided a digital output of criterial high-amplitude sucking responses. Additional equipment included a Teac 3340 tape recorder, a Kenwood (KA-3500) power amplifier, an Ads 200 loudspeaker, a Grason-Stadler (Model #1200) programmable logic board, a power supply, two relays, a counter and a Physiograph dc preamplifier. Each criterial response activated a timer on the logic board for a one second period or restarted the period. Auditory stimulation at a level of 75 ± 2 dB (A) SPL (approximately 15 dB above the background noise level caused by the ventilation system) was available whenever the timer was in an active state. By using the logic board to monitor the auditory signals on the tape recorder, it was possible to ensure that the timer was never activated in the middle of a TOT stimulus.

5. Subjects

The subjects were 96 infants, 49 males and 47 females. Mean age was 10.0 weeks (range: 7 to 13 weeks). In order to obtain 96 infants for the study, it was necessary to test 231. Subjects were excluded from this study for the following reasons: crying (33%) or falling asleep (33%) prior to shift, ceasing to suck during the course of the experiment (i.e. 2 consecutive minutes with less than 2 sucks/min) (10%), failure to maintain a minimal criterial sucking rate of 15 responses/min during the satiation period (7%), equipment failure (6%), experimenter error (6%), and miscellaneous (3%).

III. RESULTS

Figure 2 displays the mean number of high-amplitude sucking responses as a function of minutes and experimental groups. For purposes of statistical comparison, we examined each subject's rate of sucking during five intervals: baseline minute, third minute before shift, average of minutes 1 and 2 before shift, average of minutes 1 and 2 after shift, and average of all 4 minutes after shift. Difference scores were then calculated for each subject for each of the following rate comparisons: (1) acquisition of the sucking response -- third minute before shift less baseline; (2) satiation -- third minute before shift less the average of the last 2 minutes before shift; (3) release from satiation -- average of first 2 minutes after shift less the average of the last 2 minutes before shift; (4) release from satiation for the full 4 minutes -- average of 4 minutes after shift less the average of the last 2 minutes before shift.

Insert Figure 2 about here

In each of Groups III, IV, and V, half of the subjects were tested on one stimulus pair and half of the subjects on another. For each of these groups, Randomization tests for independent samples (Siegel, 1956) were used to determine whether the data from the two kinds of stimulus pairs could be pooled for further analysis. Since no significant differences between stimulus pairs emerged for any of these groups, the data were combined for further statistical treatment.

As is usually the case in studies employing the HAS procedure, subjects in all sessions acquired the conditioned high-amplitude sucking response and satiated to the first stimulus prior to shift. An indication of the mean change in response rate during the postshift period for each of the 6 groups is provided in Table 2. Randomization tests for independent samples (Siegel, 1956) were employed to assess performance during the postshift periods. Postshift performance of each of the experimental groups (II, III, IV, V and VI) was compared to that of the control group (I) for both the first 2-minute and the full 4-minute periods. These tests indicated that the only reliable ($p < .05$, 1-tailed) differences occurred between the control group and two of the Within category groups (II and VI) for both the first 2-minute and the full 4-minute periods. Neither of the two Between category groups (III and V) nor the other Within category group (IV) performed reliably differently than the control group. Although the mean change in response rate after shift was somewhat smaller for Group II than for Group VI, subsequent comparisons of these two groups by means of Randomization tests for independent samples, indicated that no reliable differences existed between them for either the first 2-minute or full 4-minute periods. Thus, as was the case for adult subjects, infants were capable of

Insert Table 2 about here

discriminating differences in the relative onset of two events. Moreover, the pattern of the discrimination data suggests the presence of three perceptual categories along this temporal continuum. However, the regions of highest discriminability observed in these infants apparently differ somewhat from our initial expectations based on the adult discrimination data.

IV. DISCUSSION

The overall results of the present study are generally consistent with our predictions based on the temporal order hypothesis of voicing perception. We have shown that infants are capable of discriminating differences in temporal order information in nonspeech signals having speech-like properties. The pattern of results indicates the presence of three well-defined perceptual categories along this temporal continuum, corresponding to leading, simultaneous and lagging events. Thus, in general, these findings provide additional support for the claim that the underlying basis of the perception of VOT in stop consonants reflects a basic limitation of the auditory system to response to differences in temporal order at stimulus onset. Therefore, the auditory system of young infants may be predisposed, in some sense, to respond to salient and well-defined properties of acoustic signals that represent the acoustic correlates of the distinctive features of speech. One such salient acoustic property appears to be the relative timing of events at stimulus onset, corresponding, in the case of voicing perception, to the temporal ordering of laryngeal and supralaryngeal events, a nearly universal property of all languages.

Although the major findings of the present study demonstrate that young infants can discriminate relatively small differences (i.e. 30 msec.) in temporal order information, the specific details of the results differ somewhat from those anticipated at the outset. Specifically, we predicted that the infants would be able to discriminate only the stimulus contrasts that were selected from opposite sides of either the -20 or +20 msec. boundary, the value assumed to represent the threshold for temporal order in adults. However, the present results indicated that infants discriminate only the -70/-40 msec. lead and the +40/+70 msec. lag contrasts, stimulus pairings that were initially assumed to represent "within category" comparisons. These findings indicate that infants' sensitivity to temporal order information is shifted slightly toward larger stimulus values on this test continuum. It is unlikely that these results are due to some artifact in the specific stimulus contrasts employed or details of the HAS measurement procedure since the shifts in discrimination occurred for both lead and lag contrasts. Moreover, these shifts were displaced in opposite directions in each case toward larger stimulus differences in temporal order. Nevertheless, it should be noted that the small discrepancy between the adult and infant discrimination data could be simply a consequence of the degree of imprecision that is present in the HAS procedure itself. Discrimination data collected in this paradigm does not permit an exact specification of the infants' sensitivity or threshold. Rather, these discrimination measures provide only a rough indication of the range over which large differences in sensitivity might be observed. Thus, the exact values obtained in any HAS discrimination study must be interpreted with care and direct comparisons between adults and infants made with some caution.[2]

The present investigation was undertaken not only to determine infants' responsiveness to temporal order information in nonspeech signals but also to examine whether temporal order information might serve as the underlying basis for the perception of VOT in speech stimuli. The previous findings of Pisoni (1977) with adults indicated a close correspondence between identification and discrimination of temporal order information in nonspeech signals and suggested a possible account of the perception of speech signals differing in VOT. However, the present results revealed a slight divergence, at least for infants, in the precise location of the region of highest discriminability for the nonspeech stimuli. While we would want to interpret this discrepancy cautiously, the results raise the possibility that temporal order per se may not be the only property that young infants respond to in discriminating VOT. As mentioned earlier, Stevens and Klatt (1974) have suggested that infants could be responding to the presence or absence of an F1 transition and not VOT. Although Lisker (1975) has questioned the importance claimed for this particular acoustic cue in controlling adults' perception of voicing differences, it may be the case that the presence or absence of a rapid spectrum change at onset serves as one of several salient properties that infants initially respond to in discriminating stop consonants. We might speculate further that in the course of perceptual development, the F1 transition information is combined in some way with other acoustic cues such as those related to processing temporal order information and that these complex or integrated cues gradually assume a larger and larger role in controlling the perception of voicing as the child's perceptual system develops. It should be noted, however, that any account of VOT perception based on the F1 transition cue is incomplete since it can only be invoked to deal with the discrimination of differences in the lag region of the stimulus continuum where the duration of the F1 transition varies inversely with VOT.[3]

With regard to accounting for the cross-language data on infants' perception of VOT, it may be necessary to assume that infants first respond to speech signals on the basis of the sensory or psychophysical properties of the stimuli without any subsequent phonetic coding or interpretation. Experience in the language-learning environment would enable infants to utilize other acoustic attributes that might be prominent in phonetic environments defined by the phonological constraints of the specific language. Differential weighting might then be assigned to these acoustic cues according to their salience in marking a distinctive contrast in the language or particular dialect. Thus, a change in the relative weightings of the acoustic cues for a particular phonetic contrast could shift the region of sensitivity along some selected stimulus continuum. According to this view, infants' reliance on a common set of psychophysical properties could be responsible for the apparent universality of VOT discrimination by infants from different language-learning environments. Differences in the relative weights assigned to the various acoustic attributes to voicing could account for the cross-language differences observed in adult speakers. Questions surrounding perceptual tuning by environmental input and a more detailed discussion of the developmental course of speech perception in infants are taken up in a recent chapter by Aslin and Pisoni (1978).

An alternative to the psychophysical account summarized above is one that assumes that the infant's discrimination of VOT is, in fact, based on some form of phonetic coding or interpretation of the stimuli as speech, a view first proposed by Eimas _et_ _al_. (1971). Given the current procedures available for studying speech perception in infants, it is extremely difficult to determine whether an infant's perceptual behavior is controlled entirely by the psychophysical or phonetic properties of the stimuli. Moreover, there is little known at this time about how these two levels of perceptual analysis interact during the course of perceptual development. One promising avenue currently open to investigators is to search for correspondences in discrimination between speech and comparable nonspeech signals. When such correspondence can be found, they would strongly imply some sensory or psychophysical basis to the perception of a particular set of acoustic correlates to a distinctive feature. For example, in the earlier study of Jusczyk _et_ _al_. (1977) on the discrimination of rise-time by infants, evidence was found with nonspeech stimuli indicating that infants can discriminate differences in tempo of frequency change. This result was subsequently verified for speech stimuli by Hillenbrand, Minifie and Edwards (1977) who reported that infants can discriminate the differences between [ba] and [wa]. Thus, one could account for these results by a common underlying factor involving the detection of rate of frequency change.

While one may still be forced into accepting at least some type of phonetic coding account for certain aspects of the infant's perceptual behavior, it may be difficult to reconcile this position with the results of recent comparative studies that have examined the perception of speech signals by animals (Kuhl & Miller, 1975, 1978). In the absence of alternative proposals, the most parsimonious explanation for the parallels observed in the perception of speech signals by animals and humans is one that also assumes a common underlying basis for the two sets of results in terms of some general psychophysical process. At the present time, there is strong evidence that both humans and chinchillas respond in somewhat similar ways to VOT, a temporal contrast. However, it remains to be seen in future work if similar evidence can be adduced for other phonetic contrasts that have well-defined acoustic properties. A psychophysically based explanation of the infant's ability to discriminate the acoustic correlates of place or manner may be somewhat more difficult to develop as the comparable nonspeech experiments examining the possible underlying perceptual properties that define these categories have not yet been conducted (see Walley & Aslin, 1979). Nevertheless, an important goal of future research will be to specify more precisely the sensory and perceptual correlates of the distinctive features in speech and how the abilities to perceive these salient properties develop in the young infant.

In summary, the present study has demonstrated that young infants can discriminate differences in temporal order information in nonspeech signals. The pattern of results suggests the presence of three well-defined perceptual categories corresponding to leading, simultaneous and lagging temporal events. Although the overall results of this study were similar to earlier work obtained with adults, several differences were observed in the precise location of the perceptual categories that could be inferred from the infant discrimination data. Despite these differences, the main findings provide some additional support for the hypothesis that the perception of VOT, one of the major cues to voicing in stop consonants, involves the perception of the relative temporal order of the component events at stimulus onset.

REFERENCES

Aslin, R.N., Hennessey, B., Pisoni, D.B. and Perey, A.J., (1979), "Individual infant's discrimination of voice onset time: Evidence for three modes of voicing". Paper presented at the Biennial Meeting of the Society for Research in Child Development (San Francisco, California).

Aslin, R.N. and Pisoni, D.B., (1978), "Some developmental processes in speech perception". Paper presented at N.I.C.H.D. conference on "Child Phonology: Perception, Production and Deviation", (Bethesda, Maryland).

Abramson, A. and Lisker, L., (1967), "Discriminability along the voicing continuum: Cross-language tests". *Proceedings of the 6th International Congress of Phonetic Sciences*.

Cutting, J.E., Rosner, B.S. and Foard, C.F., (1976), "Perceptual categories for musiclike sounds: Implications for theories of speech perception". Quart. J. Exp. Psychol., *28*, 361-378.

Cutting, J.E. and Rosner, B.S. - Categories and boundaries in speech and music. (1974), "Percept. Psychophys.", *16*, 564-571.

Eimas, P.D., (1978), "Developmental aspects of speech perception" in *Handbook of Sensory Physiology*: *Perception*, edited by R. Held, H. Leibowitz and H.L. Teuber (New York: Springer-Verlag).

Eimas, P.D., Siqueland, E.R., Jusczyk, P. and Vigorito, J., (1971), "Speech perception in infants". Science, *171*, 303-306.

Hillenbrand, J., Minifie, F.D. and Edwards, T.J., (1977), "Tempo of frequency change as a cue in speech sound discrimination by infants". Paper presented at the Biennial Meeting of the Society for Research in Child Development, New Orleans.

Hirsch, I.J., (1959), "Auditory perception of temporal order". J. Acoust. Soc. Am., 31, 759-767.

Hirsch, I.J. and Sherrick, C.E., (1961), "Perceived order in different sense modalities". J. Exp. Psychol., 62, 423-432.

Jakobson, R., Fant, G. and Halle, M., (1952), "Preliminaries to Speech Analysis", (Cambridge, Mass.: M.I.T. Press).

Jusczyk, P.W., Rosner, B.S., Cutting, J.E., Foard, C.F. and Smith, L.B., (1977), "Categorical perception of nonspeech sounds by 2-month old infants". Percept. Psychophys., 21, 50-54.

Kewley-Port, D., (1976), "A complex-tone generating program" in Research on Speech Perception: Progress Report #3 (Department of Psychology, Indiana University, Bloomington, Indiana).

Kuhl, P. and Miller, J.D., (1978), "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar - plosive consonants". Science, 190, 69-72.

Kuhl, P.K. and Miller, J.D., (1978), "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli". J. Acoust. Soc. Am., 63, 905-917.

Lasky, R.E., Syrdal-Lasky, A. and Klein, R.E., (1975), "VOT discrimination by four to six and a half month old infants from Spanish environments". J. Exp. Child. Psychol., 20, 213-225.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. and Studdert-Kennedy, M., (1967), "Perception of the Speech Code". Psychol. Rev., 74, 431-461.

Liberman, A.M., DeLattre, P.C. and Cooper, F.S., (1958), "Some cues for the distinction between voiced and voiceless stops in initial position". Lang. Speech, 1, 153-167.

Lieberman, P., (1976), "Phonetic features and physiology: a reappraisal". J. Phonetics, 4, 91-112.

Lisker, L., (1975), "Is it VOT or a first-formant transition detector?" J. Acoust. Soc. Am., 57, 1547-1551.

Lisker, L. and Abramson, A., (1964), "A cross language study of voicing in initial stops: Acoustical measurements". Word, 20, 384-422.

Lisker, L. and Abramson, A., (1967), "The voicing dimension: Some experiments in comparative phonetics". Proceedings of the 6th International Congress of Phonetic Sciences.

Miller, J.D., Engebretson, A.M., Spenner, B.F. and Cox, J.R., (1977), "Preliminary analyses of speech sounds with a digital model of the ear". J. Acoust. Soc. Am., 62, S1, 13.

Miller, J.D., Wier, L., Pastore, R., Kelly, W. and Dooling, K., (1976), "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception". J. Acoust. Soc. Am., 60, 410-417.

Morse, P. and Snowdon, C., (1975), "An investigation of categorical speech discrimination by rhesus monkeys". Percept. Psychophys., 17, 9-16.

Pastore, R.E., (1976), "Categorical perception: A critical re-evaluation", in Hearing and Davis: Essays Honoring Hallowell Davis, edited by S.K. Hirsch, D.H. Eldredge, I.J. Hirsch and S.R. Silverman (Washington University: St. Louis), pp. 253-264.

Pastore, R.E., Ahroon, W.A., Buffuto, K.J., Friedman, C.J., Puleo, J.S. and Fink, E.A., (1977), "Common factor model of categorical perception". J. Exp. Psychol.: Human Percep. Perf., 4, 686-696.

Pisoni, D.B., (1977), "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops". J. Acoust. Soc. Am., 61, 1352-1361.

Searle, C.L., Jacobson, J.Z., Rayment, S.G. and Dockendorff, D., (1979), "Phoneme recognition based on human audition". J. Acoust. Soc. Am., 65,

Siegel, S., (1950), "Nonparametric statistics for the behavioral sciences", New York: McGraw-Hill.

Sinnott, J., Beecher, M., Moody, D. and Stebbins, W., (1976), "Speech sound discrimination by humans and monkeys". J. Acoust. Soc. Am., 55, 653-659.

Siqueland, E.R. and DeLucia, C.A., (1969), "Visual reinforcement of non-nutritive sucking in human infants". Science, 165, 1144-1146.

Stevens, K.N., (1972), "The quantal nature of speech", in Human Communication: A unified view, edited by E.E. David, Jr., and P.B. Denes (New York: McGraw-Hill).

Stevens, K.N. and Klatt, D.H., (1974), "Role of formant transitions in the voiced-voiceless distinction for stops". J. Acoust. Soc. Am., 55, 653-659.

Streeter, L.A., (1976), "Language perception of 2-month old infants shows effects of both innate mechanisms and experience". Nature, 259, 39-41.

Summerfield, Q.S. and Haggard, M., (1977), "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants". J. Acoust. Soc. Am., 62, 435-448.

Walley, A. and Aslin, R.N., (1979), "Infants' discrimination of full and partial cues to place of articulation in stop consonants".

Paper to be presented at the 97th meeting of the ASA, Cambridge, Mass.

Waters, R.S. and Wilson, W.A., (1976), "Speech perception by rhesus monkeys: The voicing distinction in synthesized labial and velar stop consonants". Percept. Psychophys., 19, 285-289.

Wood, C.C., (1976), "Discriminability, response bias, and phoneme categories in discriminations of voice onset time". J. Acoust. Soc. Am., 60, 1381-1389.

FOOTNOTES

1. This research was supported, in part by an NSERC grant (A0282) to the first author and by NIH grants NS-12179-04 and HD-11915-01 and NIMH grant MH-2407-04 to Indiana University. We also wish to acknowledge our appreciation to the staff and administrators of the Grace Maternity Hospital for their support and assistance. The final paper was written while the second author was a Guggenheim Fellow at the Research Laboratory of Electronics, M.I.T. An earlier version of this paper was presented at the Biennial Meeting of the Society for Research in Child Development in San Francisco, California on March 17, 1979.

2. The use of more refined measures such as those employed by Aslin, Hennessey, Pisoni and Perey (1979) may permit a more exact specification of the infant's threshold.

3. It is worth noting here that discrimination performance in the lag region of the VOT continuum tends to be better than that for the lead region (e.g. Abramson & Lisker, 1967; Aslin & Pisoni, 1978). We might speculate that the addition of the F1 transition cue to the lag region, but not the lead region, may help to account for the better discriminability of contrasts in the lag region.

Table 1

Design and Breakdown of Experimental Groups

Group	Type of Contrast	Stimulus Pairings
I	Control (e.g. Lead_1 vs Lead_1)	-70 vs -70, -40 vs -40, etc.
II	Within (Lead_1 vs Lead_2)	-70 vs -40
III	Between (Lead vs. Simul.)	-40 vs -10 (8 Ss) -30 vs 0 (8 Ss)
IV	Within (Simul_1 vs Simul_2)	-20 vs +10 (8 Ss) -10 vs +20 (8 Ss)
V	Between (Simul. vs Lag)	0 vs +30 (8 Ss) +10 vs +40 (8 Ss)
VI	Within (Lag_1 vs Lag_2)	+40 vs +70

Table 2

Mean Change in Response Rate After Shift

	Release from Satiation (Minutes After Shift)	
Group	First 2	Full 4
I (Control)	-0.03	-0.34
II (Within-Lead)	6.00*	6.28*
III (Between-Lead/Simul.)	3.63	3.73
IV (Within-Simul.)	-0.88	-2.02
V (Between-Lag/Simul.)	2.28	2.55
VI (Within-Lag)	11.50*	8.97*

*Indicates a reliable difference ($p<.05$ or better) when compared to the performance of control subjects for the same period.

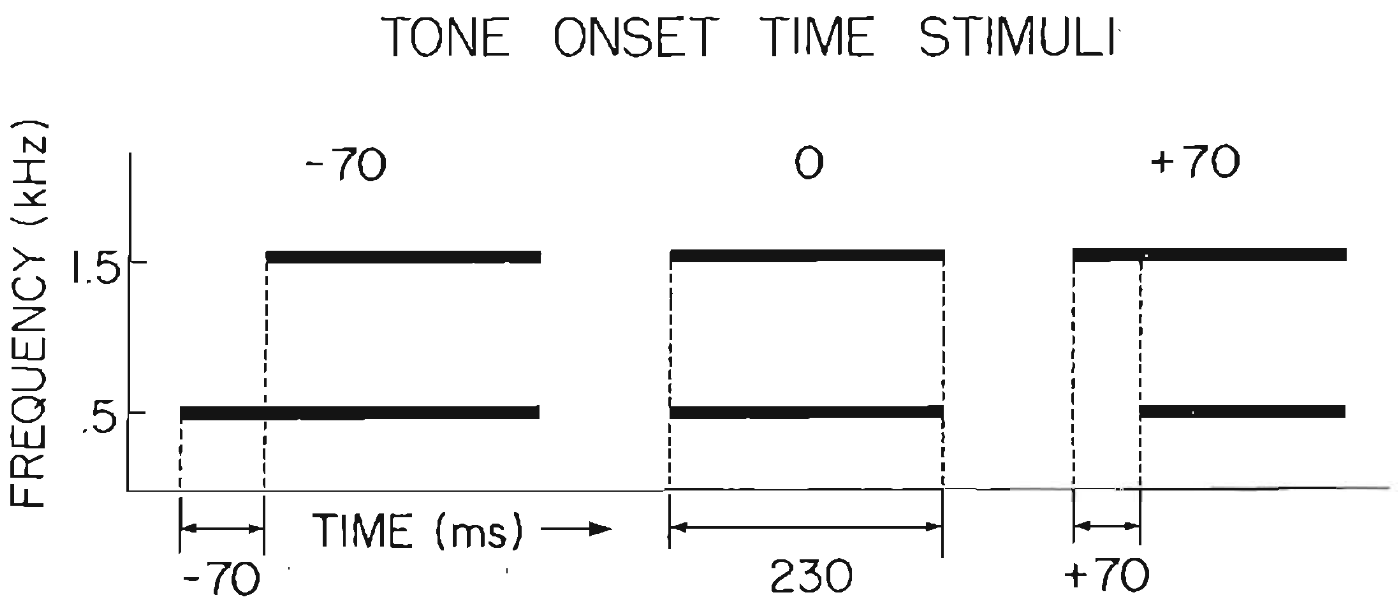


Figure 1. Schematic representations of three stimuli differing in relative onset time: leading (-70 msec.), simultaneous (0 msec.), and lagging (+70 msec.).

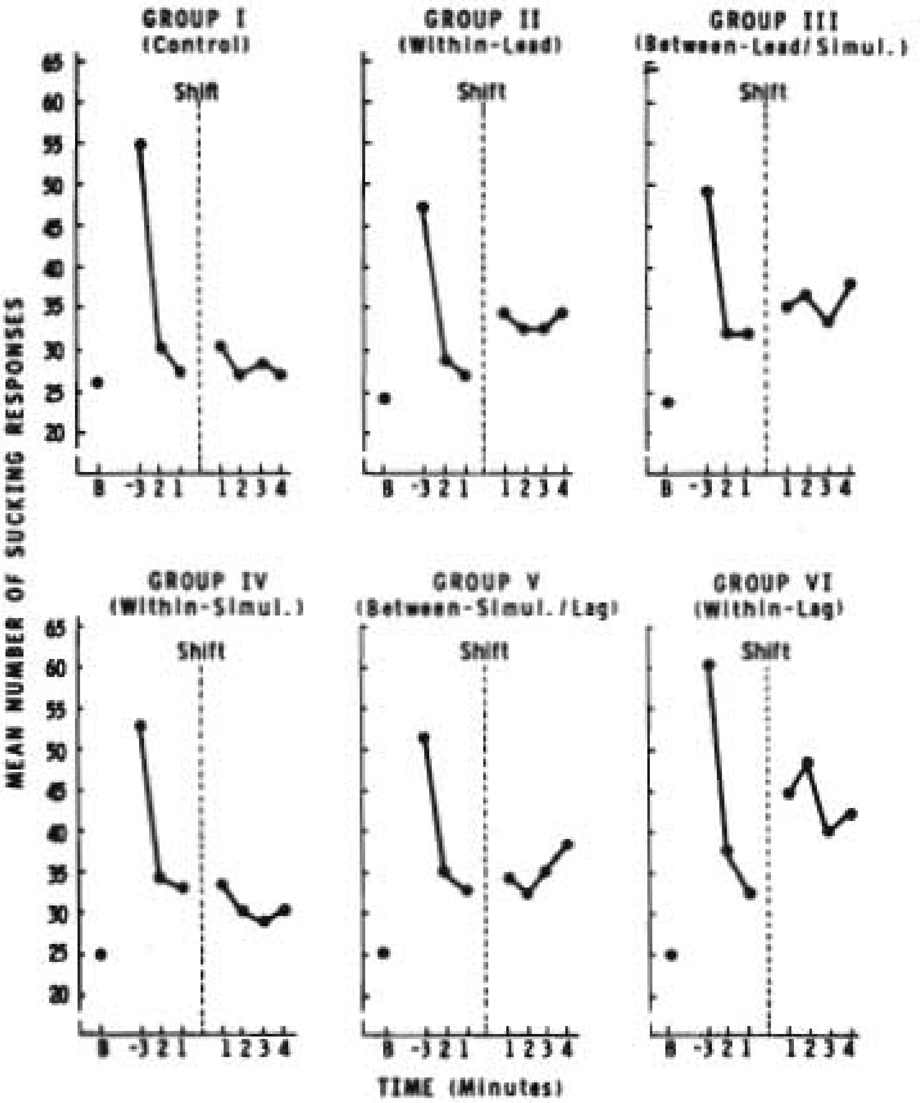


Figure 2. Mean number of high amplitude sucking responses as a function of time and experimental group. Time is measured with reference to the moment of the stimulus shift, marked by the vertical dashed line. The baseline rate of sucking is indicated by the letter "B".

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 5 (1978-79) Indiana University]

Infants' Discrimination of Full and Partial Cues to Place of Articulation in Stop Consonants

Amanda Walley

Indiana University

Running head: Infants' Discrimination of Place of Articulation

Abstract

It has recently been proposed that the shape of the onset spectrum provides contextually invariant information about place of stop consonant articulation for consonant-vowel syllables and that these primary spectral attributes underlie the infant's ability to discriminate place differences. According to this view, contextually variable formant transitions constitute only secondary, learned cues to place of articulation. Prelinguistic infants are, therefore, assumed to be incapable of discriminating place differences in two-formant stimuli which supposedly lack invariant spectral attributes. Our analysis revealed, however, that the two-formant labial and velar CV stimuli are spectrally very similar to their full-formant counterparts. Therefore, if spectral shape does cue place of articulation, infants should actually discriminate some of the two-formant place contrasts, in addition to the full-formant ones. An operant head-turning paradigm was employed to test this hypothesis with 6-9 month-old infants. The results were consistent with the proposal that spectral cues mediate place perception, but the finding that infants were able to discriminate two-formant stimuli at the same time renders the distinction between primary vs. secondary cues of little theoretical value for understanding the perception of place of articulation.

Infants' Discrimination of Full and Partial Cues to Place of Articulation

A great deal of research has focused on describing the role of the various acoustic attributes of the speech signal in determining the perception of the phonetic distinctions employed by language. Several investigators have, for example, attempted to specify the acoustic cues which enable adults to differentially identify the voiceless /p,t,k/ and voiced /b,d,g/ stop consonants that differ in place of articulation (labial, alveolar, velar, respectively). Apparently, variations in the starting frequencies and directions of the second- and third-formant transitions are an important cue to place of articulation (Liberman, Cooper, Shankweiler and Studdert-Kennedy, 1967). A schematic representation of this cue is shown in Figure 1. However, it has been demonstrated that the spectrum of the burst of acoustic energy at consonantal release also provides information about place of articulation (e.g., Cooper, Delattre, Liberman, Borst and Gerstman, 1952; Liberman, Delattre and Cooper, 1952; Schatz, 1954). Furthermore, it has

----

Insert Figure 1 about here

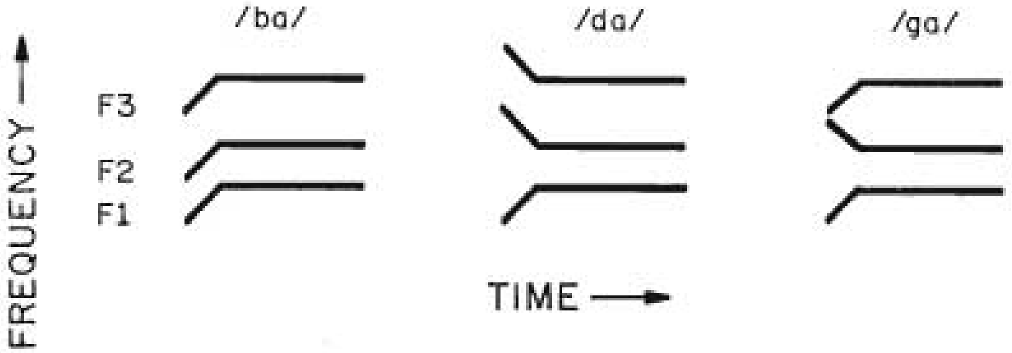


Figure 1. Schematic representation of the formant (F) transition patterns associated with labial, alveolar and velar voiced stop consonants followed by the vowel /a/.

been suggested that the direction of the rapid spectrum change immediately following consonantal release is a more appropriate way of characterizing the acoustic differences underlying the perception of different places of stop consonant articulation (Stevens, 1975; Stevens and Blumstein, 1975).

Although all of these acoustic properties have been implicated in the perception of place of articulation, attempts to state unequivocally which serves as the primary cue for this phonetic feature have been complicated by the fact that all of these acoustic features may vary for a given place of articulation in different vowel contexts. For example, the second-formant transition falls following consonantal release in /da/, but rises in /di/ and has a different starting frequency in these two syllables (but see Delattre, Liberman and Cooper, 1955). The failure to find a set of absolute, invariant properties with which to characterize place of articulation is representative of one of the most pervasive problems in speech perception research; the speech signal typically exhibits, as a function of phonetic context, speaking rate, speaker identity and stress, a great deal of acoustic variability and there is a marked lack of any simple correspondence between acoustic and phonetic segments (Liberman et al., 1967). Researchers are thus confronted with the task of accounting for the observed constancy of the phonetic percept and its derivation from the soundstream.

It has been maintained by some investigators that starting frequency and transition direction (particularly of the second-formant) do, despite their contextual variability, constitute the major cue for place of articulation (e.g., Liberman et al.,1967). As Stevens and Blumstein (1978) have pointed out, the invariance of the phonetic percept could, in keeping with this claim, only result from the interpretation of this cue in a manner that is different for each environment in which a given stop occurs. In order to explain how such contextually determined interpretation might be achieved, a number of theories have found it necessary to view the speech perception process as an active one which, at some level beyond peripheral auditory processing, uses rather extensive knowledge of the phonological structure of language and perhaps even higher levels of linguistic knowledge to impose structure on the acoustic elements of the speech signal (e.g., Chomsky and Miller, 1963; Chomsky and Halle, 1968; Liberman et al., 1967; Stevens and Halle, 1967; Stevens and House , 1972). There does indeed exist a large body of empirical evidence to support this notion. If linguistic knowledge is actually a prerequisite for the perception of speech, this might logically require that experience in speech perception and/or production is necessary to perceive, for example, contextual variations in acoustic features as belonging to the same phonetic categories. This implication has been challenged, however, in light of the results of studies of infant speech perception (Eimas, 1975).

Several studies have shown that prelinguistic infants perceive speech sounds in much the same way that adults do. The first of these studies revealed, for example, that one- and four-month-old infants are able to discriminate speech stimuli varying along the dimension of voice onset time (VOT) better when these stimuli are from different (i.e., voiced vs. voiceless) adult phonemic categories than when they are from the same categories (Eimas, Siqueland, Jusczyk and Vigorito, 1971). VOT is defined for word-initial stop consonants as the interval between the onset of the release burst and the onset of laryngeal pulsing (Lisker and Abramson, 1964). Like adults then, infants appear to perceive VOT information in a categorical manner. With only very limited exposure to the numerous phonetic distinctions employed by various languages and with no experience in the consistent articulation of these distinctions, infants are somehow able to sort certain acoustic variations into the appropriate adult phonemic categories and ignore within-category variations. While active theories of speech perception might still account for this ability of infants by positing innate knowledge of phonological rules, Eimas (1975) has proposed that the infant data would be accomodated better by simply assuming the existence of a linguistic feature detector system. He has argued that the human auditory system might be endowed with feature detectors which are sensitive to the restricted ranges of acoustic information that signal phonetic features and thus be predisposed to perceive certain speech stimuli in a linguistically relevant way (Eimas, 1975).

Recently, Pisoni (1977) has provided convincing evidence for an alternative, psychophysical explanation of the categorical perception of the VOT continuum. He demonstrated that when the relative onset times of two- component tones are varied, adults perceive such variations in these nonspeech stimuli categorically. Stimuli with onset differences greater than 20 msec are perceived as successive events, those with onset differences less than 20 msec as simultaneous ones. Similar results have been obtained recently with infants (Jusczyk, Pisoni, Walley and Murray, Note 1). The category boundary values observed by Pisoni (1977) correspond very closely to the loci along the VOT continuum of the three voicing categories found earlier by Lisker and Abramson (1964) across a wide variety of languages. This correspondence suggests that the categorical perception of VOT information may simply reflect an inherent limitation of the auditory system to resolve the temporal relation between two events at stimulus onset (i.e., between the onset of the release burst and that of laryngeal pulsing) and not the operation of a mechanism that is specialized for speech processing. Presumably, certain languages have, in the manner that Stevens (1972) has suggested, exploited this particular property of the auditory system as one way of representing voicing distinctions. It may be more appropriate, therefore, to attribute the categorical perception of the VOT continuum to acoustic feature detectors, rather than to phonetic feature (i.e., voicing) detectors per se. This interpretation is supported by other demonstrations that infants and adults do, in fact, perceive complex nonspeech stimuli categorically (e.g., Cutting and Rosner, 1974; Jusczyk, Rosner, Cutting, Foard and Smith, 1977; Miller, Wier, Pastore, Kelley and Dooling, 1976). In addition, the results of several selective adaptation studies indicate that larger phonetic boundary shifts are produced when adapting and test stimuli share acoustic, and not just phonetic, features (e.g., Ainsworth, 1975; Diehl, 1975; Ganong, 1975; Sawusch, 1977). Finally, it has been shown that the chinchilla exhibits categorization of stop consonants that is similar to that of humans, even though the chinchilla's perception is obviously not mediated by the phonological system of any natural language (Kuhl and Miller, 1975; Kuhl and Miller, 1978).

The importance of these putative acoustic feature detectors is that they provide the basis for a mode of perception which accounts, in part, for the invariance of the voicing percept and thus the child's ability to acquire voicing as a phonemic contrast. Although it seems reasonable to assume that a psychophysical basis for the categorical perception of variations in place of articulation may also exist, such a basis is less obvious in view of the greater contextual variability of the various hypothesized cues to this feature (see, however, Pisoni, Note 2). It is precisely in order to identify such a psychoacoustic basis that it is essential to determine what the primary cue for place of articulation in speech is. Yet even if some property of the auditory system could be invoked to account for the categorical perception of this primary cue, this would not necessarily provide a complete explanation of the observed perceptual invariance of place of articulation in stops. Categorical perception can explain perceptual constancy within phonetic categories in the face of variations along a particular acoustic dimension, but how is it that different acoustic features (such as formant transitions vs. release bursts) in the same syllable position and large contextual variations in a given acoustic feature (e.g., in transition starting frequency and direction) can be perceived similarly? For example, in a CV syllable, alveolar place of articulation can be cued by a falling transition in one vowel context and by a rising one in another vowel context. In the appropriate vowel context, the same rising transition can, on the other hand, signal labial place of articulation. Thus, while some psychophysical account could be offered for the categorical perception results dealing with variations in transition direction, this would not, in itself, provide a complete explanation of perception of place of articulation; i.e., it would still remain to be determined how, in different vowel contexts, one category on a continuum of transition variations can signal different places of articulation and how different categories can, in the appropriate vowel contexts, signal the same place of articulation.

Although a few studies have addressed the question of whether or not the perceptual equivalence of different and/or contextually diverse acoustic features exists for infants (Eilers, 1977; Fodor, Garrett and Brill, 1975; Kuhl, 1976), investigations of infant speech perception have, for the most part, only examined discrimination of stimuli varying along a particular acoustic dimension (for a review, see Eimas, 1978). It is not at all clear from these investigations that the sort of complete form of perceptual constancy just described for adults is to be found in infants. This is an important consideration when evaluating the account of perception of place of articulation offered by Stevens and

Blumstein (1978), since it is exactly this sort of ability (perceptual constancy) which they imply infants possess and which they attempt to explain. Stevens and Blumstein object to the view that only after learning to organize contextually verse and variable acoustic features into their appropriate adult phonemic categories does the child come to perceive place of articulation distinctions. Instead, they assume that some innate mechanism must mediate such organization because several studies have shown that infants do, in fact, discriminate place of articulation differences (Bush and Williams, 1978; Eimas, 1974; Leavitt, Note 5; Leavitt, Brown, Morse and Graham, 1976; Miller and Morse, 1975; Moffitt, 1971; Morse, 1972). It must be emphasized here that these earlier studies have merely shown that infants are capable of discriminating place of articulation differences in stimuli varying along a particular acoustic dimension. They have not demonstrated that infants perceive syllables such as /da/ and /di/ as being in some sense similar and support or evidence for perceptual constancy cannot, therefore, be inferred from previous studies of infants' perception of place of articulation. Until it is shown that infants are able to sort different and/or contextually variant acoustic features into their appropriate adult phonemic categories, theories which require experience in the perception and/or production of these features are not, as Eimas (1975) contends and Stevens and Blumstein implicitly assume, invalidated on the basis of the current data from infant speech perception research (but see Pisoni, 1977,for other criticisms of these theories).

While it has not yet been demonstrated that infants have any initial basis for recognizing that contextual variations in acoustic features belong to certain phonetic categories, Stevens and Blumstein are reluctant to abandon the notion that some invariant property exists in the acoustic correlates of each particular place of articulation category. They argue that even though various context-dependent features, such as starting frequency and direction of formant transitions and release bursts, are separately observable in a spectrogram, the auditory system does not necessarily "process" these features independently of one another. Instead, they argue that the auditory system integrates these features in such a way that the gross spectral properties associated with each place of articulation category provide the acoustic invariance which must, in their opinion, underlie the constancy of the phonetic percept and mediate infant perception. The search for invariant acoustic correlates of phonetic features and thus for a means of automatic, passive recognition of phonetic distinctions represents a noteable digression from the proposal that speech perception proceeds primarily via the active operations entailed in analysis-by-synthesis (Stevens and Halle, 1967; Stevens and House, 1972).

Stevens and Blumstein's (1978) assertion that there are distinctive and context-independent acoustic properties associated with different places of stop consonant articulation derives from both theoretically based expectations about the gross shape of the short-term spectrum sampled at consonantal release and from preliminary spectral analyses of natural speech. Labials are characterized by a diffuse- falling spectrum, alveolars by a diffuse-rising spectrum and velars by a prominent mid-frequency spectral peak (see Figure 2). These putatively invariant acoustic cues for place of articulation - location and diffuseness of spectral energy at stimulus onset - are, of course, very similar to the compact vs. diffuse and grave vs. acute features originally proposed by Jakobson, Fant and Halle (1952). Onset spectra characteristics are determined by the burst spectrum and the initial portions of the formant transitions at voicing onset. The same spectral shapes can be obtained for stimuli containing only formant transitions and no burst, but these shapes are enhanced by the presence of the burst. Stimuli with only the burst cue do not yield these distinctive spectral shapes. Because Stevens and Blumstein found that only those stimuli with distinctive spectral

----

Insert Figure 2 about here

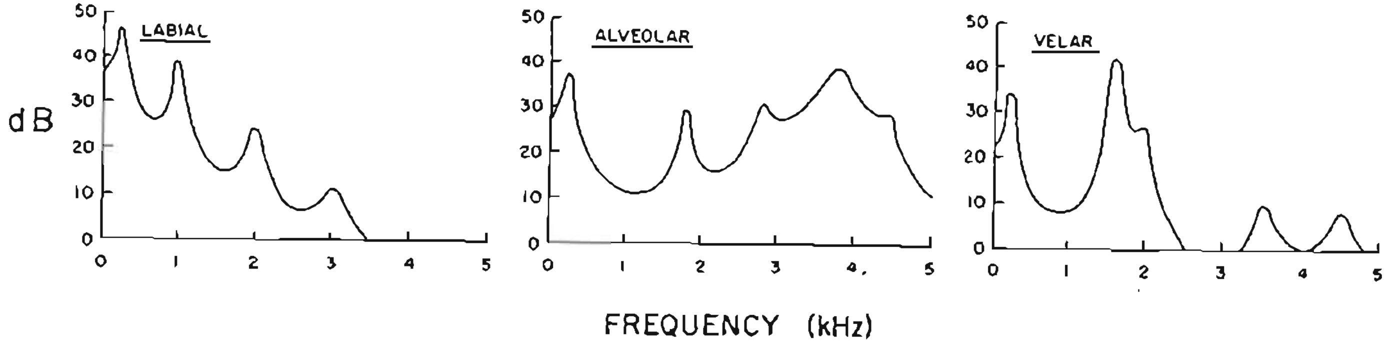


Figure 2. Representation of the context-independent spectra associated with the labial, alveolar and velar places of articulation with release bursts present (after Stevens and Blumstein, 1978).

characteristics were identified consistently by adults according to place of articulation, they proposed that the auditory system also performs a short-term spectral analysis at stimulus onset for a stop consonant. According to this view, formant transitions are not the primary cue to place of articulation. Rather, identification of this phonetic feature is achieved through the operation of detectors which, at the peripheral stage of auditory processing, are sensitive to invariant properties of the onset spectrum.

There are several problems with Stevens and Blumstein's theory. First, it may be noted that this theory is not, strictly speaking, adequate to account for the perception of natural speech tokens of /b,d,g/ (Kewley-Port, Note 3). Natural exemplars of alveolar and velar places of articulation typically produce a burst cue of relatively greater and longer frication than that of a labial, and the second- and third-formant transitions of the velar are somewhat longer in duration than those of the labial and alveolar. Therefore, because the time window selected by Stevens and Blumstein for sampling the spectrum of the stimulus waveform is of the same duration (a half-width of 26 msec) for all three places of articulation, a sample for a natural labial stimulus would, therefore, include substantial portions of the following steady-state vowel. The shape of the spectrum of this stimulus would not be invariant with respect to vowel context, but would depend rather heavily on the properties of the following vowel. Stevens and Blumstein have circumvented this problem by constructing their synthetic stimuli in such a way that only the burst cue and the formant transitions are, in fact, sampled by a time window of predetermined size. They are correct in stating that, for their particular stimuli, the gross shape of the spectrum for a stop consonant is determined independently of vowel context. Using natural speech tokens, they have met with some success in verifying their theory (Blumstein and Stevens, Note 4).

Stevens and Blumstein (1978) postulate that the auditory system is endowed with feature detectors which are sensitive to the context-independent spectral properties of stop consonants and that these detectors account for the infant's ability to discriminate stimuli with different places of articulation - particularly when these stimuli contain both transitions and bursts. This assertion, which is based, in part, on the work of Bush and Williams (1978), deserves some comment. Bush and Williams examined infants' discrimination of alveolar and velar place of articulation using two sets of three-formant stimuli - one containing both bursts and transitions (full cues), and the other containing transitions only (partial cues). From their results, these investigators inferred greater discriminability for the full cue stimuli even though the evidence for greater discriminability did not reach an acceptable level of statistical significance. Furthermore, Bush and Williams failed to include in their study a crucial control condition in which the burst cue was simply added to the training stimulus (e.g., /da/ vs. /da/ + burst). Thus, what they interpreted as greater discriminability for the full cue stimuli could be attributable to the greater salience of the burst itself, rather than to the integration of burst and transition cues (Aslin and Pisoni, Note 5). Therefore, not only do these results fail to conclusively support Stevens and Blumstein's theory, but they might just as well be interpreted as indicating that it is primarily the formant transitions which cue place of articulation.

The claim that context-independent properties are associated with and mediate perception of a given place of articulation might be challenged on the basis of yet another finding. It has been shown that adults are able to differentially identify two-formant stimuli with respect to place of articulation (Cooper et al., 1952; Delattre et al., 1955; Liberman et al., 1952), although two-formant stimuli do not, so Stevens and Blumstein report (1978), yield spectra of the distinctive, contextually invariant shapes which purportedly underlie the perception of this phonetic feature. Stevens and Blumstein agree that, in two-formant stimuli, only the second-formant transition can signal differences in place of articulation and proceed to explain the adult's ability to use this context-dependent cue in terms of the co-occurrence of primary, invariant and secondary, context-dependent features in the full formant stimuli. Because adults have learned these co-occurrences through repeated exposure to them in the linguistic environment, they can, in the absence or distortion of the primary attributes of the stimulus, use the secondary cue of starting frequency and direction of the second-formant transition to identify place of articulation.

By proposing that formant transitions constitute a secondary and learned cue to place of articulation, Stevens and Blumstein's theory makes several predictions about infants' perception of place of articulation. First, because infants have had little exposure to speech and presumably have not yet learned which variations in formant transitions co-occur with the invariant spectral properties of a particular place of articulation, they should be unable to discriminate formant transition variations that are not accompanied by different spectral properties. For example, for either two or three-formant stimuli, infants should not be able to discriminate the two exemplars of the alveolar stop in /da/ and /di/ on the basis of the information provided by the transitions alone. As mentioned earlier, it is, in fact, an assumption of Stevens and Blumstein's theory that infants do not perceive differences in such contextual variations within a given phonetic category. However, according to the view that formant transitions provide the primary cue to place of articulation, infants might be expected to discriminate these stimuli, since the second-formant transition falls in /da/, but rises in /di/. Only if the results of such investigation merited the interpretation of the perceptual equivalence of contextual variations in formant transitions, could Stevens and Blumstein's theory be considered viable. Unfortunately, this prediction is virtually impossible to test with the current discrimination paradigms that are used to study infant perception because such a test would be confounded by differences in the acoustic properties of the vowel. Another test of the sort of perceptual constancy which Stevens and Blumstein seem to assume would be to somehow determine how infants generalize a learned discrimination, such as /ba/ vs. /da/, to the syllables /ab/ and /ad/. If the formant transitions provide the basis for the initial discrimination, /ba/ ought to be generalized to /ad/, and /da/ to /ab/, since the transitions are rising and falling, respectively, in these pairs of syllables. Stevens and Blumstein have, on the other hand, expressed the idea that the spectrum sampled at the vowel offset of a VC syllable should exhibit the same properties as the onset spectrum for a given place of articulation. Within the framework of their theory, one would predict then that /ba/ and /da/ should be generalized to their appropriate phonetic categories.

A more immediate test of Stevens and Blumstein's theory would be available if, as they report, two-formant stimuli differing in place of articulation did not yield the distinctive, contextually invariant spectra of their full-formant counterparts. According to Stevens and Blumstein's theory, formant transitions provide only a secondary, learned cue to place of articulation and infants should not, therefore, be able to use this cue to discriminate two-formant stimuli differing in place of articulation in the way that adults do. If, on the other hand, formant transitions do constitute the major cue for this phonetic feature, infants should be able to discriminate differences in two formant stimuli. Some evidence for the utilization of single transition cues for place of articulation discrimination in infants has already been provided. With three-formant stimuli, it has been shown that infants can discriminate labial vs. velar place of articulation (Moffitt, 1971; Morse, 1972; Leavitt, Note 6; Leavitt _et al._, 1976) and alveolar vs. velar place of articulation (Miller and Morse, 1976). In terms of formant transitions, the first distinction is differentiated solely by that of the second-formant and the second distinction by that of the third-formant. Stevens and Blumstein would, of course, argue that it is the distinctive shape of the onset spectra of these three-formant stimuli which underlies the infant's discrimination. However, Eimas (1974) found that infants presented with two-formant stimuli can also discriminate labial vs. alveolar stops which are differentiated solely by the second-formant transition. It cannot be asserted that discrimination here is mediated by a divergence in spectral shape if stimuli containing only the first two-formants do not possess the distinctive and invariant spectra that three-formant stimuli do. This suggests then that it is the second-formant transition which provides the basis for the infant's discrimination, although, according to Stevens and Blumstein, infants should not be able to use this cue in discrimination.

Because the two accounts of the perception of place of articulation under consideration here appear to make different predictions about infants' discrimination of two-formant stimuli, the present experiment was originally intended to provide a strong test of these two accounts by determining whether or not infants do discriminate place differences in two-formant stimuli. After the two-formant stimuli for the present experiment were constructed and their onset spectra analyzed, it was noted, however, that the onset spectra of the labial and velar stimuli were, contrary to Stevens and Blumstein's report, very similar in shape to those of the full-formant stimuli (see Figure 3). Of course, the two-formant alveolar stimulus differed from the full formant one (an obvious consequence of removing the upper formants) and was actually very similar to the two-formant velar stimulus. The discrepancy between Stevens and Blumstein's claim and our initial findings concerning the onset spectra of two-formant stimuli is perhaps a result of the fact that the two-formant stimuli used in previous studies were constructed on a more primitive speech synthesizer than those used for the present experiment and were, therefore, of marginal quality. Furthermore, subject feedback was used to modify our two-formant stimuli such that adult subjects were eventually able to achieve over 90% correct identification for these stimuli. Presumably, these particular two-formant stimuli represent rather good perceptual approximations to the syllables /ba/, /da/ and /ga/ and possess many of the acoustic attributes of their full-formant counterparts.

Insert Figure 3 about here

Our preliminary findings concerning the onset spectra of the two-formant stimuli clearly renders Stevens and

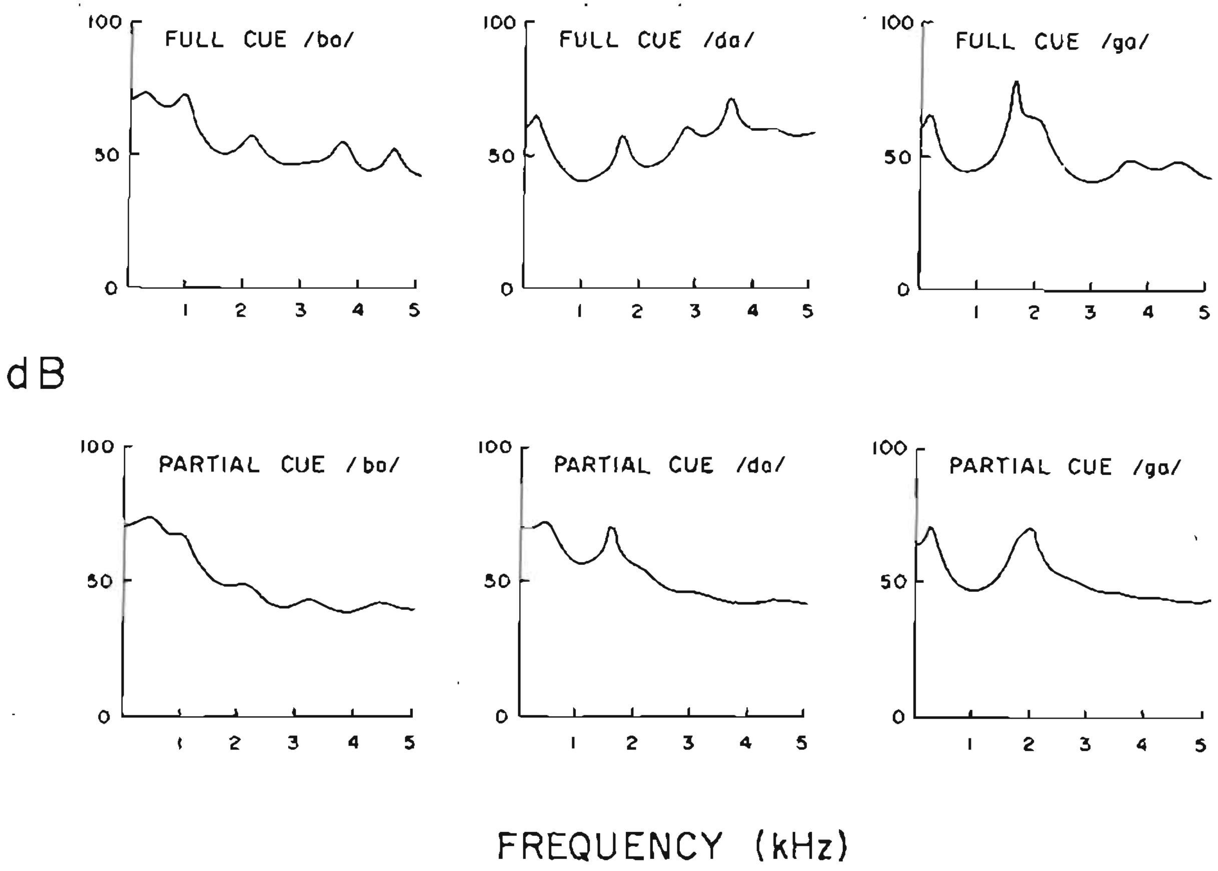


Figure 3. Onset spectra for the full and partial cue labial, alveolar and velar stimuli used in the present experiment.

Blumstein's proposal that formant transitions constitute secondary, learned cues to place of articulation of little theoretical value for understanding the perception of place of articulation; it may well be that spectral attributes mediate place perception, but since the putatively primary, invariant spectral cues typically occur even in so-called degraded (i.e., two-formant) stimuli, there seems to be no logical necessity for an infant to learn to use contextually diverse formant transition starting frequency and direction as an additional cue to place of articulation (at least in the case of the labial vs. alveolar and labial vs. velar two-formant contrasts). Moreover, if, as Stevens and Blumstein maintain, it is differential sensitivity to spectral shape that mediates perception of place of articulation differences, then infants should indeed be able to discriminate the labial vs. alveolar and labial vs. velar contrasts in the two-formant stimuli. Thus, Stevens and Blumstein's theory actually makes the same predictions about infants' discrimination of two-formant stimuli as does the the notion that formant transitions provide the primary cue to this feature.

It might still be possible, however, to differentiate the two accounts on the basis of their predictions concerning infants' discrimination of the alveolar vs. velar two-formant contrast. Because the spectra for these two stimuli are very similar, infants should not, according to Stevens and Blumstein's theory, be able to make this discrimination. If formant transition starting frequency and direction cue place of articulation, infants might be expected to discriminate this contrast - depending, of course, on their frequency resolution.

It is of interest to note here that Kuhl and Miller (1975) failed to find any evidence for the utilization of single transition cues by chinchillas for the the discrimination of place of articulation. Whereas chinchillas are able to discriminate the labial vs. alveolar contrast, which is differentiated, in three-formant stimuli, by both the second- and third-formant transitions, they are not able to discriminate labial vs. velar or alveolar vs. velar stimuli. These two latter contrasts involve only single formant transition differences. This outcome is significant to the extent that the discriminative abilities of the chinchilla (which has an auditory system comparable to the human one) are usually very similar to those of humans. This similarity is typically interpreted as evidence against the existence of perceptual abilities in the human which are specific to speech processing. Kuhl and Miller have argued that the chinchilla's failure to make discriminations involving only single formant differences is simply the consequence of lower frequency resolution -- i.e., that this failure stems from a difference in auditory, and not a lack of phonetic, processing of the speech signal. This explanation seems to be a plausible one since there do exist anatomical and physiological differences between the chinchilla and the human auditory systems. The basilar membrane of the chinchilla is, for example, shorter than that of man. Kuhl and Miller's interpretation of their results seems to be more in keeping with the notion that formant starting frequency and transition direction provide the primary cue for place of articulation; assuming that the auditory systems of chinchillas and humans do function similarly with respect to the perception of speech stimuli, one would still expect, within the framework of Stevens and Blumstein's theory, that chinchillas would be able to discriminate labial vs. velar and alveolar vs. velar stimuli; i.e., they ought to have the same access as humans do to the gross spectral properties of three-formant stimuli.

The present experiment employed a variation of the operant head-turning (OHT) procedure first developed by Eilers, Wilson and Moore (1977) to assess infants' discrimination of two- and full-formant (+burst) stimuli differing in place of articulation. In this way we hoped to pursue the issue of which acoustic features mediate the perception of this phonetic feature.

An attempt was first made to verify that infants can discriminate the three place of articulation contrasts, /ba/ vs. /da/, /ba/ vs. /ga/ and /da/ vs /ga/, when they are specified by full formant stimuli containing both bursts and formant transitions (i.e., full cues). Although Bush and Williams' earlier work (1978) suggested that infants can discriminate these contrasts, these investigators only examined discrimination of the /da/ vs. /ga/ contrast. The stimuli for this part of the present experiment duplicated, as closely as was possible, those employed by Stevens and Blumstein (1978).

In the event that infants discriminate place of articulation differences in the full cue stimuli, we wished to address the question of which acoustic features of the full cue stimuli mediate this discrimination performance. If formant starting frequency and transition direction provide the primary cues for place of articulation, infants would be expected to discriminate place of articulation differences in two-formant or partial cue stimuli, as well as those in the full cue stimuli. Eimas (1974) has provided some evidence to this effect, but he studied only discrimination of the labial vs. alveolar contrast. The present experiment tested infants' discrimination of all three of the possible contrasts between labial, alveolar and velar places of articulation and employed two-formant stimuli which were derived from the full-formant stimuli. In light of the preliminary finding of the present experiment that two of the two-formant stimuli are characterized by the distinctive, invariant onset spectra comparable to their full-formant counterparts, infants would, according to Stevens and Blumstein's theory, also be expected to discriminate the labial vs. alveolar and the labial vs. velar two-formant contrasts, but not the alveolar vs. velar one.

Method

Subjects.

A total of 62 six- to nine-month-old infants, with no known hearing disorder, participated in the present experiment. Subjects were obtained through a system of infant subject solicitation already establshed in the Psychology Department at Indiana University. Parents of prospective subjects previously identified in the birth announcements of the local newspaper were sent a letter explaining the purpose and procedure of the study one week prior to testing and were later contacted by telephone so that appointments could be scheduled for those wishing to participate in the experiment. Infants were typically retested on successive days and parents were paid $3.00 on completion of each experimental session.

Stimuli.

A set of five-formant stimuli with both burst and transition cues (full cue) and a set of two-formant stimuli with the transition cue only (partial cue), the members of which corresponded perceptually to the syllables /ba/, /da/ and /ga/, were generated on a modified version of Klatt's digital speech synthesizer. This program is currently implemented on a PDP-11/05 in the Speech Perception Laboratory at Indiana University (Kewley-Port, 1978). These stimuli were modelled after stimuli constructed by Stevens and Blumstein (1978) on an earlier version of the same synthesizer (Klatt, 1972). Except where otherwise noted, all of the parameter values used in the synthesis of the stimuli for the present experiment were identical to those used by Stevens and Blumstein.

Each of the stimuli was synthesized at a 10 KHz sampling rate and low passed at 4.8 KHz on output through a 12-bit D-A converter. The duration of voicing for all stimuli was 260 msec. The excitation source for the vowel began abruptly, such that the first glottal pulse coincided with the beginning of the formant transitions. The amplitude was constant for 255 msec and fell to 0 dB over the last 35 msec. The fundamental-frequency contour began at 103 Hz, rose to 125 Hz in 40 msec, fell first to 94 Hz in 180 msec and then to 125 Hz in 40 msec.

Full Cue Stimuli. The three full cue stimuli - each synthesized with the digital resonators connected in series - were modelled after the best exemplars of each place of articulation category (i.e., after stimulus 1, 8 and 12) in the Stevens and Blumstein (1978) full cue /ba,da,ga/ series. The center frequencies and bandwidths of the formants were appropriate for the steady-state vowel /a/ and were set at 720 (50), 1240 (70), 2500 (110), 3600 (170) and 4500 (250) Hz, for F1, F2, F3, F4 and F5, respectively. All of the stimuli had the same starting frequency (220 Hz) for F1, but had varying transition durations - 20, 35 and 45 msec for /ba/, /da/ and /ga/, respectively. The starting frequencies of F2 and F3 were 900 Hz and 2000 Hz for /ba/, 1700 Hz and 2800 Hz for /da/, 1640 Hz and 2100 Hz for /ga/. The transition durations of F2 and F3 were 45 msec. F4 and F5 were steady-state formants and thus had no formant transitions.

Consistent with the construction of the Stevens and Blumstein stimuli, the peak of the burst spectrum for a given stimulus was located at the starting frequency of a particular formant of the following vowel. This procedure was adopted, according to Stevens and Blumstein, in order to satisify the condition of continuity between release burst and vowel formants - a condition which arises from the fact that, in natural speech, the same vocal tract resonances are excited first by noise and later, at voicing onset, by the glottal source although with different relative amplitudes. A burst, consisting then of a single resonance peak, was located at 900 Hz (BW=70) for the labial stimulus, at 3600 Hz (BW=170) for the alveolar stimulus and at 1610 Hz (BW=70) for the velar stimulus. The bursts were produced by exciting these formants with 5 msec of random noise, which began 5, 10 and 15 msec prior to voicing onset for /ba/, /da/ and /ga/, respectively. Total stimulus duration for the labial, alveolar and velar stimuli was, therefore, 265, 270 and 275 msec, respectively.

Because of differences in the two versions of the Klatt synthesizer used for stimulus construction, the amplitude of the burst relative to the vowel for the stimuli of the present experiment could not be set with the same synthesis parameter values of the Stevens and Blumstein stimuli. Linear prediction analysis was, therefore, employed to match spectral sections of the original Stevens and Blumstein stimuli with the waveforms of the stimuli of the present experiment and to thus determine the appropriate amplitudes for the frication of the bursts. The original Stevens and Blumstein stimuli were digitized from audiotapes provided by Professor Stevens and a linear prediction analysis program was used for the spectral analysis (see Kewley-Port, Note 7). Because it was observed that there was greater energy present in the region of F4 and F5 relative to F1 in the steady-state vowel of the present stimuli, the tilt of the vowel was adjusted slightly by modifying the parameters controlling the bandwidth of the glottal antiresonator and the lip radiation characteristic (see Klatt, 1977).

Partial Cue Stimuli. Each of the partial cue stimuli was synthesized with the digital resonators connected in parallel. The parameter values of the full cue stimuli were followed as closely as possible in constructing the corresponding labial, alveolar and velar partial cue stimuli. Thus, the starting frequencies of F1 and F2 for the partial cue /ba/ and /da/ and the steady-state values for the formants and bandwidths of F1 and F2 for all of the partial cue stimuli were identical to those of their full cue counterparts. Several modifications were, however, necessary in order for the partial cue stimuli to be identified consistently according to place of articulation.

Feedback from naive adult listeners, who were required to achieve 90% correct identification of the stimuli, was used to determine the optimal parameter values for the partial cue stimuli. Specifically, the starting frequency of F2 in the velar stimulus was extended to 1940 Hz. In addition, the transition duration of F1 for the alveolar stimulus was reduced to 20 msec. Finally, the formant amplitudes, which could be manipulated independently of formant frequency, since these stimuli were constructed using the parallel mode of the synthesizer were altered so that the relative amplitudes of F1 and F2 in the full and partial cue stimuli were the same. The spectral analysis program was employed to set the amplitudes and achieve the best spectral matches. The partial cue stimuli were each 260 msec in duration. Their spectro-temporal specifications, together with those of the full cue stimuli, are shown graphically in Figure 4.

Insert Figure 4 about here

Design. Subjects were tested individually in several experimental sessions conducted on successive days. Each subject was randomly assigned to one of 12 experimental groups. Groups 1, 2 and 3 received the contrasts /ba/ vs. /da/, /ba/ vs. /ga/, /da/ vs. /ga/, respectively, with first the full cue stimuli and then the partial cue stimuli. Groups 4, 5 and 6 were presented with the same contrasts, respectively, but with the stimulus order reversed. Groups 7-9 were first tested on one of the partial cue contrasts and then on the corresponding full cue contrast. Groups 10-12 were tested on the same contrasts , respectively, but with the stimulus order reversed.

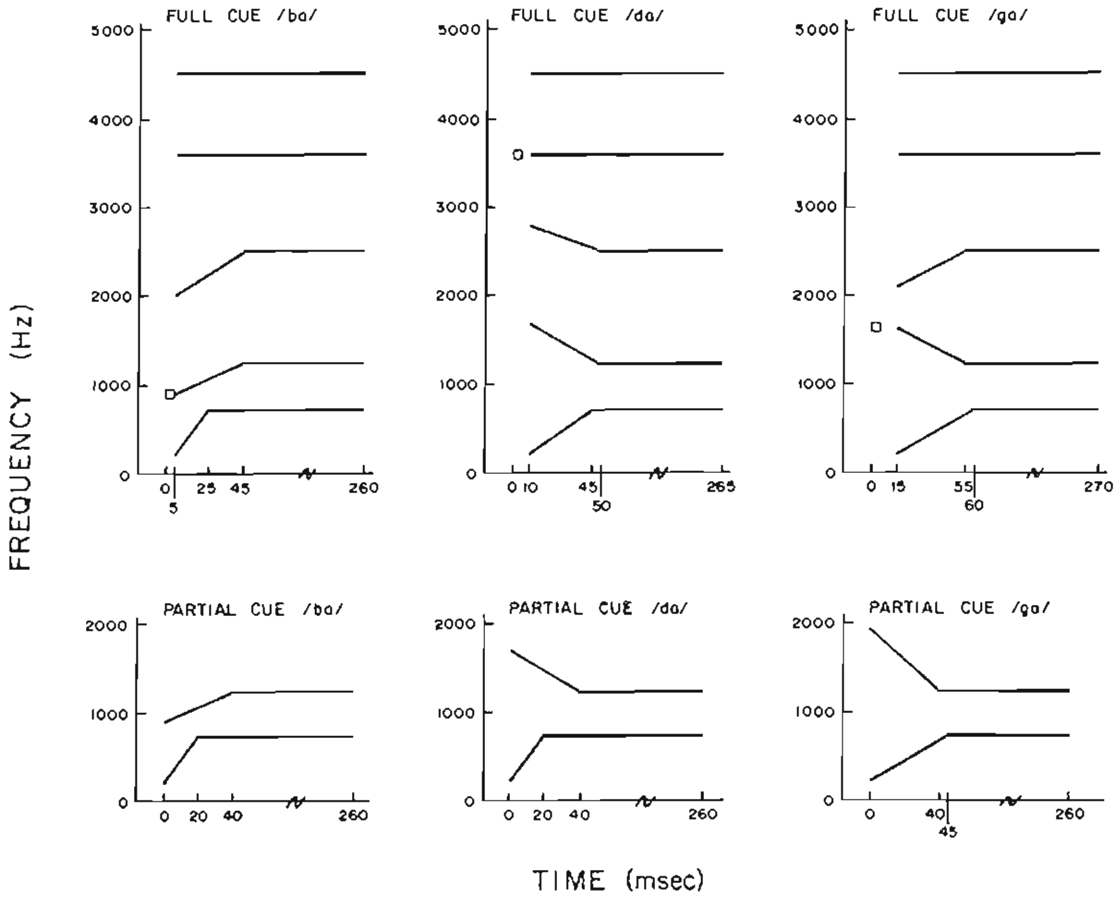


Figure 4. Spectro-temporal specifications for the full and partial cue stimuli used in the present experiment.

Procedure. An operant head-turning procedure was employed in the present experiment to measure discrimination of the contrasts presented. Inside a sound-attenuated IAC booth, an assistant attracted the gaze of the infant who was seated on the parent's lap. A repeating background stimulus (the second member of a given stimulus contrast) was presented at a rate of once per second to the infant. Both the assistant and the parent listened to masking music over head- phones during the entire experimental session to prevent any biasing of the infant's responses. An outline of the experimental setting is shown in Figure 5.

Insert Figure 5 about here

In the initial shaping phase of the procedure, the infant learned to make a directional head-turn in anticipation of the presentation of a visual stimulus (a mechanical toy), which served as a reinforcer, whenever a change in the speech stimulus occurred. An experimental trial was initiated by the experimenter, who was located outside the booth and monitored the infant's behavior via closed circuit TV. If the infant responded to a stimulus change with a head-turn toward the speaker from which the speech stimuli were delivered, a visual reinforcer was presented for 3 seconds. At all other times, the reinforcer, located inside a smoked plexi-glass container, was not visible to the infant.

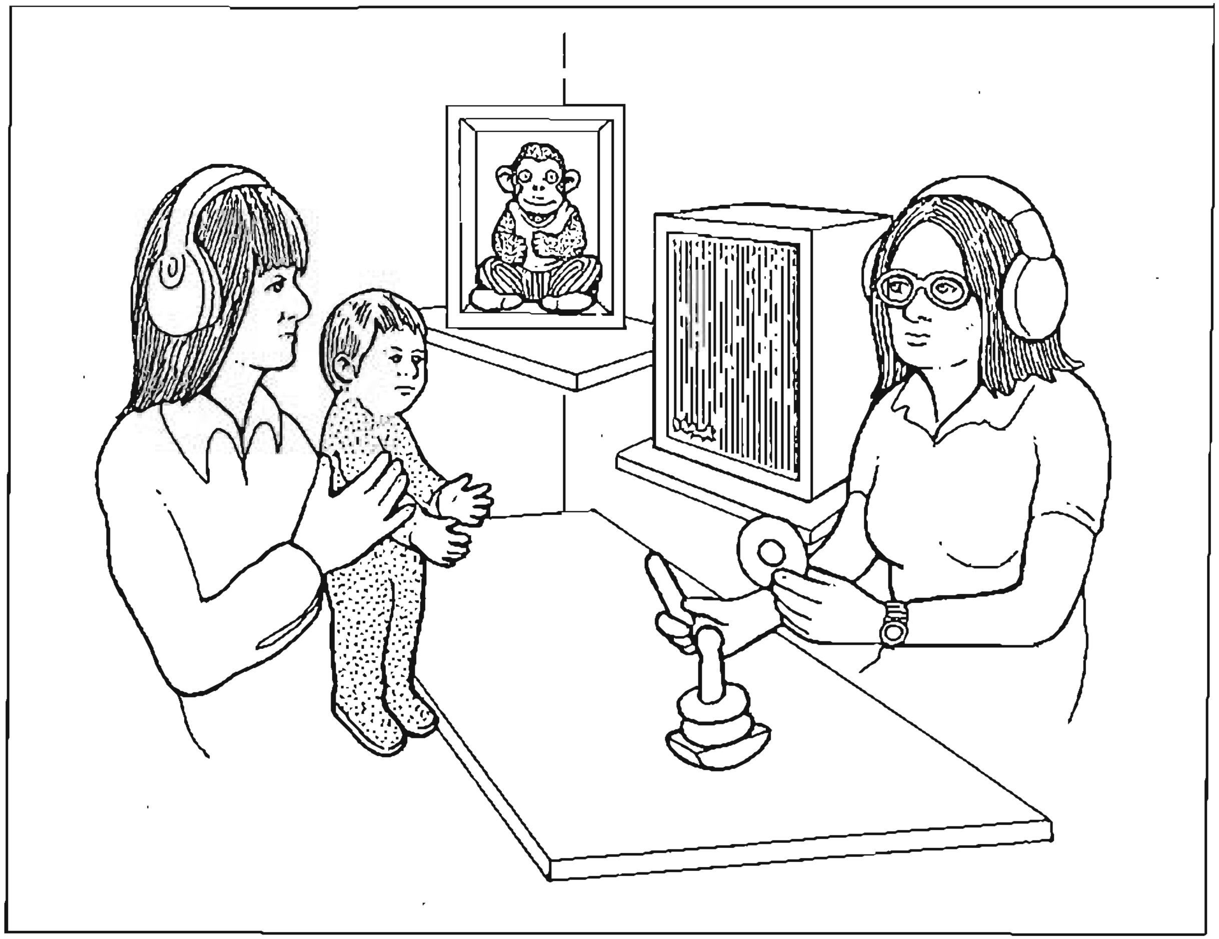


Figure 5. Interior of sound-attentuated booth during the testing session employing the operant head-turning paradigm (after Aslin *et al*., Note 8).

On initial shaping trials, if the infant did not respond to the sound change, the visual reinforcer was presented to elicit a head-turn coincident with the speech stimulus change. The target stimulus was initially set at 10 dB above the level of the background stimulus to facilitate shaping of the head-turn response. This difference was attenuated in 5 dB steps until the infant responded consistently to a 0 dB difference between the background and target stimuli, at which point an immediate transition to the testing phase of the procedure was made.

In the testing phase of the procedure, one third of the trials initiated by the experimenter consisted of no change from the background to the target stimulus (i.e., control trials), while the other two thirds consisted of change trials (i.e., experimental trials). During this phase of testing, the experimenter wore headphones and was blind to the specific stimulus conditions and reinforcement contingencies on any given trial. A tone presented over the headphones indicated the occurrence of a four second scoring interval within which the experimenter recorded any head-turns. A head-turn on an experimental trial constituted a correct response and a head-turn on a control trial an incorrect response. This procedure thus provided a measure of the proportion of correct responses after a speech stimulus change and incorrect responses (false alarms) after no stimulus change. Infants were required to achieve 80% correct responding on a minimum of 5 experimental and 5 control blocked trials in order to successfully complete testing on a contrast. When this was achieved, they were further required to meet the same criterion in an additional block of 10 trials before being tested on the same contrast with the other set of stimuli (i.e., with either the full or partial cue stimuli). All experimental events -- stimulus presentations, response recording and presentations of the visual reinforcer -- were controlled on-line in real-time by a PDP-11 computer.

Results

Of the 62 infants participating in the experiment, 17 failed to show any evidence of acquiring the head-turn response after 2 sessions and were dropped from the present study. Table 1 shows the proportion of the remaining 45 subjects who met the discrimination criterion in at least one block of 10 trials for the first (either full or partial cue) condition in which they were tested and for the particular place contrast, collapsed across target-background stimulus order, which they received.

Insert Table 1 about here

Table 1

Proportion of Subjects

Meeting Discrimination Criterion

	Cue Type			
Place Contrast	Full Cue	n	Partial Cue	n
ba-da	.86	7	.17	6
ba-ga	.31	16	.67	3
da-ga	.09	11	.50	2
Total	.35	34	.36	11

First it may be noted that a certain proportion of the subjects did meet this criterion for the full cue contrasts (.35) and the partial cue contrasts (.36). The difference between these proportions is not significant (z=.06, p>.05), which may be indicative of the equal discriminability of the full and partial cue contrasts. While the proportion of subjects that discriminated the full cue contrasts is somewhat low overall, this outcome is due primarily to the small number of subjects who completed testing on the /d/ vs. /g/ contrast. Second, calculation of a 99% confidence interval for the mean difference in the number of trials required to meet criterion by subjects tested first in the full and partial conditions revealed that this difference is probably no greater than 6 trials. This may be taken as another indication that the two classes of stimuli are equally discriminable.

A comparison of the proportion of subjects in groups 1-6 meeting criterion in the full cue condition shown in Table 1 indicates that a greater proportion of subjects were able to discriminate the /b/ vs. /d/ contrast than the /b/ vs. /g/ contrast (z=3.35, p<.01) and the /d/ vs. /g/ contrast (z=2.50, p<.05), but that there is no difference in the proportion of subjects able to discriminate the /b/ vs. /g/ and the /d/ vs. /g/ contrasts (z=1.29, p>.05). Although this latter comparison does not reach statistical significance, there does appear to be a tendency for a greater proportion of subjects to discriminate the /b/ vs. /g/ contrast than the /d/ vs. /g/ contrast. The number of subjects in groups 7-12 is too small, especially for the /d/ vs. /g/ contrast, to make the same comparisons for subjects who received the partial cue contrasts first.

Because both of the hypothesized cues to place of articulation are present to some extent (i.e., in two of the three possible contrasts) in the full and partial cue stimuli, then presumably, if infants use one cue to discriminate one class of stimuli, they can use the same cue to discriminate the other class; i.e., if infants are able to discriminate a contrast for one class of stimuli, it would be expected that they also discriminate the same contrast for the other class of stimuli. In order to determine whether such a relationship between performance for the full and partial cue stimuli exists, one would ideally want to examine the within-subject performance data for the two classes of stimuli. A relationship between performance on the two classes of stimuli would not, of course, indicate which cue was being used, but such a dependency would, nevertheless, be expected.

Unfortunately, it was not always possible to obtain a measure of discrimination performance for each infant in both the full cue and partial cue conditions since some infants failed to reach criterion on two blocks of trials in the first condition and because several of the infants who did meet this criterion could not be rescheduled for further testing. In order, therefore, to obtain some indication of the relatedness of the two classes of stimuli, the following tests were carried out: the proportion of subjects meeting criterion on the partial cue stimuli, given that they completed testing on the full cue stimuli was compared to the proportion of subjects who were first tested on the partial cue stimuli and met criterion using a chi-square test of independence; likewise, the proportion of subjects meeting criterion on the full cue stimuli, given that they completed testing on the partial cue stimuli, was compared to the proportion of subjects who were first tested on the full cue stimuli and met criterion. Although the results of these tests were not statistically significant ($F(1)=2.10$, $p>.05$; $F(1)=1.74$, $p>.05$), the data do, nevertheless, suggest that performance (i.e., success or failure to meet criterion) in one condition may be related to performance in the other; 5 of the 7 infants who met criterion in the full cue condition and who were subsequently tested on the same partial cue contrast, did meet criterion on at least one block of trials in this second condition. The one subject who met criterion in the partial cue condition and who was tested later in the full cue condition also completed two blocks of trials in this condition.

Discussion

The results of the present experiment serve to verify previous reports that prelinguistic infants can discriminate full cue exemplars of stop consonant-vowel syllables differing in place of articulation. This particular discrimination does, nevertheless, appear to represent a difficult one for infants as evidenced by the low proportion of subjects meeting our strict discrimination criterion. Whereas this proportion is not unlike that typically observed for younger infants using the high-amplitude sucking and the heart-rate deceleration discrimination paradigms (e.g., Eimas, 1974; Miller and Morse, 1975; Morse, 1972), the OHT procedure is typically able to provide positive evidence for the discrimination of linguistic contrasts in a higher proportion of the infants tested (e.g., Aslin, Note 8; Aslin, Hennessy, Pisoni and Perey, Note 9). It might be inferred, therefore, from the results of the present study, which is the first to use the OHT procedure to extensively examine place of articulation perception, that this discrimination constitutes a relatively difficult one for infants. It should not be concluded that infants are unable to discriminate this feature, since a number of infants did meet the statistically significant criterion of 80% correct discrimination on at least 10 trials.

In a discussion of the appropriate interpretation of the failure to demonstrate the presence of various communicative skills in young children, Shatz (1974) has pointed out that, within the framework of information processing theory, children, like adults, may be viewed as limited-capacity processors and that increased task complexity may often be responsible for the failure to observe skills which are perhaps not yet firmly established, but nevertheless contained in the child's repertoire. This may explain why an ability, such as place discrimination, is readily observable in infants at one age (e.g., by means of the HAS and HR procedures, in which the cognitive demands of the task are presumably minimal), but may have difficulty manifesting itself later. Jusczyk (Note 10) has also encountered some difficulty, using the OHT procedure, in obtaining positive evidence of discrimination for this same feature in six- to nine-month-old children.

The relatively low total proportion of subjects discriminating the full cue contrasts may also be due to the small number of subjects meeting discrimination criterion for the full cue /da/ vs. /ga/ contrast and perhaps the foregoing discussion is really only applicable to this particular contrast. As the results of the present experiment indicate, the proportion of subjects discriminating /ba/ vs. /da/ is quite high and significantly greater than those discriminating the other two contrasts tested. While the proportion of subjects completing testing on the /ba/ vs. /ga/ contrast is statistically no different from the number completing /da/ vs. /ga/, there does appear to be a tendency for more subjects to discriminate the former contrast. Thus, to the extent that the proportion of subjects meeting discrimination criterion is indicative of the ease (or difficulty) of discriminability, there do seem to be differences in the discriminability of the three place contrasts. According to either the notion that spectral cues mediate place perception or the idea that formant transition starting frequency and direction do so, this general pattern of results might be anticipated since /ba/ and /da/ are the most divergent stimuli in terms of both of these putative cues. The /ba/ vs. /ga/ and /da/ vs. /ga/ contrasts can, on the other hand, be regarded as equally different. The tendency for the latter contrast to be discriminated less than the former may, therefore, be indicative of poor high frequency resolution in infants. Whereas this interpretation is consistent with Kuhl and Miller's (1975) data on the chinchilla's performance, it conflicts with several previous reports that infants of an even younger age can make this discrimination (i.e., Bush and Williams, 1978; Eimas, 1974; Miller and Morse, 1975). According to Stevens and Blumstein's theory, infants should have little difficulty making this discrimination for the full cue stimuli since the shapes of the onset spectra for the alveolar and velar stimuli are very different from one another.

The results of the present experiment also show that infants are able to discriminate place of articulation in some of the partial cue contrasts. Because there was no significant difference between the proportion of subjects meeting criterion in the partial cue and the full cue conditions as a whole, it might be concluded that the two classes of stimuli are equally discriminable. This interpretation is supported by the relatively small mean difference between the two conditions in the number of trials required to meet discrimination criterion. As was true of the full cue stimuli, discrimination of place for the partial cue stimuli (if the performance of subjects tested first in the full cue condition is also considered) was observed particularly for the /ba/ vs. /da/ and the /ba/ vs. /ga/ contrasts -- a result which would also be predicted by both theories. Unfortunately, the number of subjects tested first in the partial cue condition was too small to actually make the same comparison that was made earlier for the full cue stimuli.

Although the results of the present experiment indicate that infants can discriminate some of the two-formant stimuli, it is still not clear whether they use the shape of the onset spectrum or the dynamic transitional cues contained in the CV syllables to do so. As Stevens and Blumstein themselves state (1978), a strong test of their theory would be to vary the two types of cues independently of one another and to ascertain which cue perception follows. The present experiment was originally intended to do just this, since it was assumed that two-formant stimuli do not contain the spectral cues of their full-formant counterparts. A preliminary spectral analysis of the two-formant stimuli revealed that this is not the case. It was subsequently proposed that the results of the /da/ vs. /ga/ partial cue condition might still serve as a way of distinguishing between the two accounts of place of articulation perception under consideration in the present paper, but in view of the apparent difficulty of the same contrast for the full cue stimuli and the small number of subjects who have as of yet been tested in the partial cue condition, it is probably too early to say anything conclusive about this contrast.

Given the findings of the present experiment concerning the shape of the onset spectra of two-formant stimuli, some other means of manipulating spectral and transitional cues independently of one another is required in order to provide a stronger test of Stevens and Blumstein's theory. Specifically, they suggest that the relative amplitudes of the formants at the onset of CV syllables might be changed to yield different gross spectral shapes without changing the formant starting frequencies. After doing so, one would then want to determine whether or not perception of place of articulation depends on attributes of the gross shape of the onset spectrum - independent of burst characteristics and formant starting frequencies. Alternatively, one might, as suggested earlier, attempt to answer this question by determining whether infants can discriminate /da/ vs. /di/ or by determining how they generalize a learned discrimination, such as /ba/ vs. /da/, to the syllables /ab/ and /ad/.

Presumably the same cue that mediates discrimination of full cue contrasts also mediates discrimination of partial cue contrasts since both of the hypothesized cues to place are present to some extent in both types of stimuli. This does not hold true, however, for the /da/ vs. /ga/ partial cue contrast because the alveolar stimulus lacks the higher frequency components of its full-formant counterpart and is very similar spectrally to the two-formant velar stimlus. One might, therefore, expect that performance in one condition would be related to or predictive of performance in the other. For example, if an infant discriminates one of the full cue contrasts, then he/she should discriminate the same two-formant contrast. The results of the chi-square test for association suggest, however, that the two conditions are independent of one another, although, as mentioned previously, there is some tendency towards association of the two conditions and perhaps with a greater number of infants tested in both conditions such a dependency might emerge more clearly from the data.

Regardless of whether or not any difference in discriminability exists between the full and partial cue stimuli or between the three partial cue contrasts, the fact remains that some of our subjects were able to complete testing to criterion on the partial cue stimuli. This important finding argues rather strongly against one of the primary assertions of Stevens and Blumstein's theory - namely, that formant transitions constitute only secondary or learned cues to place of articulation in stop consonants. It may very well be that formant transitions are only secondary, learned cues to place of articulation and that infants (and adults) use spectral cues for place perception. However, since we have shown that the partial cue stimuli possess the very same invariant spectral cues as their full cue counterparts, Stevens and Blumstein's distinction between primary vs. secondary cues does not seem to be a useful one for understanding the perception of place of articulation in stops. If it is true that contextually invariant and distinctive spectral cues generally characterize even such degraded stimuli as the two-formant ones used here, logically there would seem to be little reason for learning to use the secondary cues; i.e., what advantage would there be to learning the secondary cues, when the primary ones are always available? It would, of course, be important to examine the spectra of other two-formant stimuli containing different vowels to ensure that this observation can be generalized to other contexts. If, on the other hand, formant transition starting frequency and direction do cue place of articulation, then obviously this cue is not "learned" in any traditional sense, since we have shown that infants can, indeed, discriminate two-formant stimuli.

Definitive evidence as to whether or not those characteristics of the onset spectrum described by Stevens and Blumstein mediate place of articulation perception in infants is still lacking due to the absence of the appropriate identification or generalization data. Yet it is clear from the present experiment that the primary, innate vs. secondary, learned cue distinction entailed in Stevens and Blumstein's theory is probably incorrect and is, more generally, of little use with regard to understanding the development of speech perception abilities in young infants. In recent years it has become apparent that the traditional dichotomy between genetic and experiential factors in development is an extremely difficult one to maintain. It is generally agreed that both factors are involved in the determination of behavior and that the nature of their interaction is, therefore, the important topic to be pursued. Finally, while the prelinguistic infant's discriminative capacities for various linguistic contrasts are an important topic for study, future research must also attempt to determine what the categorization abilities of the infant and the older child are since this aspect of perception is probably the most relevant to the speech perception process.

Reference Notes

1. Jusczyk, P. W., Pisoni, D. B., Walley, A. C. and Murray, J. Discrimination of relative onset time of two-component tones by infants: Some implications for voicing perception in stop consonants. Paper presented at the Biennial Meeting of the Society for Research in Child Development in San Francisco, California on March 17, 1979.

2. Pisoni, D. B. Some Remarks on the Perception of Speech and Nonspeech Signals. Paper presented at the Ninth International Congress of Phonetic Sciences in Copenhagen, August 6-11, 1979.

3. Kewley-Port, D. Personal communication, January, 1979.

4. Blumstein, S. E. and Stevens, K. N. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. Unpublished manuscript, 1979.

5. Aslin, R. N. and Pisoni, D. B. Perception of speech and nonspeech sounds in infants. Research proposal submitted to the National Science Foundation, 1977.

6. Leavitt, L. Infant cardiac orienting response to speech and non-speech stimuli. Paper presented at the Annual Meeting of the ASHA in Las Vegas, Nevada in November, 1974.

7. Kewley-Port, D. Spectrum. In D. B. Pisoni (Ed.), Research in Speech Perception Progress Report No. 5. Bloomington: Department of Psychology, Indiana University, 1978-1979.

8. Aslin, R. N. Personal communication, September, 1979.

9. Aslin, R. N., Hennessy, B. L., Pisoni, D. B. and Perey, A. J. Individual infants' discrimination of VOT: evidence for three modes of voicing. Paper presented at the Biennial Meeting of the Society for Research in Child Development in San Fransisco, California on March 17, 1979.

10. Jusczyk, P. W. Personal communication, July, 1979.

References

Ainsworth, W. A. Mechanisms of selective feature adaptation. Perception and Psychophysics, 1977, 21 (4), 365-370.

Bush, L. & Williams, M. Discrimination by young infants of voiced stop consonants with and without release bursts. Journal of the Acoustical Society of America, 1978, 63 (4), 1223-1226.

Chomsky, N. & Halle, M. The Sound Pattern of English. New York: Harper and Row, 1968.

Chomsky, N. & Miller, G. A. Introduction to the formal analysis of natural languages. In R. D. Luce, R. Bush & E. Galanter (Eds.), Handbook of Mathematical Psychology, Vol. 2. New York: John Wiley and Sons, 1963, 269-231.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M. & Gerstman, L. J. Some experiments on the perception of synthetic speech sounds. Journal of the Acoustical Society of America, 1952, 24 (6), 597-606.

Cutting, J. E. & Rosner, B. S. Categories and boundaries in speech and music. Perception and Psychophysics, 1974, 16, 564-570.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 1955, 27 (4), 769-773.

Diehl, R. L. The effect of selective adaptation on identification of speech sounds. Perception and Psychophysics, 1975, 17, 48-52.

Eilers, R. E. Context sensitive perception of naturally produced stop and fricative consonants by infants. Journal of the Acoustical Society of America, 1977, 61 (5), 1321-1336.

Eilers, R. E., Wilson, W. R. & Moore, J. M. Developmental changes in speech discrimination in infants. Journal of Speech and Hearing Research, 1977, 20, 766-780.

Eimas, P. D. Auditory and linguistic processing of cues for place of articulation by infants. Perception and Psychophysics, 1974, 16 (3), 513-521.

Eimas, P. D. Developmental aspects of speech perception. In R. Held, H. W. Leibowitz & H.-L. Teuber (Eds.), Handbook of Sensory Physiology Vol VIII: Perception. Berlin: Springer Verlag, 1978.

Eimas, P. D. Speech perception in early infancy. In L. B. Cohen & P. Salapatek (Eds.), Infant Perception, Vol. 2. New York: Academic Press, 1975.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W. & Vigorito, J. Speech perception in infants. Science, 1971, 171, 303-306.

Fant, G. Acoustic theory of speech production. The Hague: Mouton, 1960.

Fant, G. Speech sounds and features. Cambridge, Massachusetts: MIT Press, 1973.

Fodor, J. A., Garrett, M. F. & Brill, S. L. Pi ka pu: The perception of speech sounds by prelinguistic infants. *Perception and Psychophysics*, 1975, *18*, 74-78.

Ganong, W. F. An experiment on "phonetic adaptation". *MIT Quarterly Progress Report*, 1975, *116*, 206-210.

Jakobson, R., Fant, C. G. M. & alle, M. *Preliminaries to Speech Analysis*. Technical Report No. 13 Acoustics Laboratory, Massachusetts Institute of Technology, May, 1952.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F. & Smith, L. B. Categorical perception of non-speech sounds by two-month-old infants. *Perception and Psychophysics*, 1977, *21*, 50-54.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. *Research on Speech Perception Progress Report No. 4*, Indiana University, Bloomington, Indiana, 1978.

Klatt, D. Acoustical theory of terminal analog speech synthesis. *Proceedings of the 1972 International Conference on Speech Communication and Processing*, Boston, MA, 1972.

Klatt, D. A cascade-parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis. *Journal of the Acoustical Society of America*, 1977, *61*, Suppl. 1, S68 (A).

Kuhl, P. K. Speech perception in early infancy: The acquisition of speech-sound categories. In S. K. Hirsh, D. H. Eldredge, I. J. Hirsh & S. R. Silverman (Eds.), *Hearing & Davis: Essays Honoring Hallowell Davis*. St. Louis, Mo.: Washington University Press, 1976.

Kuhl, P. K. & Miller, J. D. Speech perception in the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 1975, *190*, 69-72.

Kuhl, P. K. and Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, 1978, *63*, 905-917.

Leavitt, L. A., Brown, J. A., Morse, P. A. & Graham, F. K. Cardiac orienting and auditory discrimination in 6-week infants. *Developmental Psychology*, 1976, *12*, 514-523.

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

Liberman, A. M., Delattre, P. C. & Cooper, F. S. The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, 1952, *65*, 497-516.

Lisker, L. & Abramson, A. S. A cross language study of voicing in initial stops: Acoustical measurements. *Word*, 1964, *20*, 384-422.

Miller, C. L. & Morse, P. A. The "heart" of categorical speech discrimination in young infants. (Research Status Report No. 1) Madison: University of Wisconsin, Infant Development Laboratory, August, 1975.

Miller, J. D., Wier, C. C., Pastore, R., Kelley, W. J. & Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60 (2), 410-417.

Moffitt, A. R. Consonant cue perception by twenty-four-week-old infants. Child Development, 1971, 42, 717-731.

Morse, P. A. The discrimination of speech and nonspeech stimuli in early infancy. Journal of Experimental Child Psychology, 1972, 14, 477-492.

Pisoni, D. B. Identification and discrimination of the relative onset of two-component tones: Implications for the perception of voicing in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

Sawusch, J. Peripheral and central processes in the selective adaptation of place of articulation in stop consonants. Journal of the Acoustical Society of America, 1977, 62 (3), 738-750.

Schatz, C. The role of context in the perception of stops. Language, 1954, 30, 47-56.

Shatz, M. The relationship between cognitive processes and the development of communication skills. In W. J. Arnold & D. Levine (Eds.), Nebraska Symposium on Motivation (Vol. 25). Lincoln: University of Nebraska Press, 1977.

Stevens, K. N. The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David, Jr. & P. B. Denes (Eds.), Human Communication: A Unified view. New York: McGraw Hill, 1972.

Stevens, K. N. The potential role of property detectors in the perception of consonants. In G. Fant & M. A. A. Tatham (Eds.), Auditory Analysis and Perception of Speech. New York: Academic Press, 1975.

Stevens, K. N. & Blumstein, S. E. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 1978, 64 (5), 1358-1368.

Stevens, K. N. & Blumstein, S. E. Quantal aspects of consonant production and perception: A study of retroflex consonants. Journal of Phonetics, 1975, 3, 215-234.

Stevens, K. N. & Halle, M. Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), Models for the Perception of Speech Visual Form. Cambridge, Ma.: Academic Press, 1967.

Stevens, K. N. & House, A. S. Speech perception. In J. Tobias (Ed.), Foundations of modern auditory theory: Volume II. New York: Academic Press, 1972.

Footnotes

We thank Jerry C. Forshee and David Link for their technical assistance, Diane Kewley-Port for assisting in the speech analysis and synthesis, Wendy Crawford for assisting in the data collection process, and especially the parents and infants who participated in the study. This research was partially supported by grants from NIMH (MH-24027-05) and NIH (NS-12179-04 and HD-11915-02).

Some Experiments on Perceptual Learning of Mirror-Image Acoustic Patterns

Mary Ellen Grunke and David B. Pisoni

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Short Title: Learning of Mirror-Image Acoustic Patterns

Abstract

It is well-known that the formant transitions of stop consonants in CV and VC syllables are roughly the mirror image of each other in time. These formant motions reflect the acoustic correlates of the articulators as they move rapidly into and out of the period of stop closure. Although acoustically different, these formant transitions are perceived as similar by adults. Earlier research had suggested that mirror image acoustic patterns resembling formant transitions were not perceived as similar on a psychophysical basis when adult subjects had to scale them according to similarity. However, mirror image patterns could still have some underlying similarity which might facilitate learning, recognition and the establishment of perceptual constancy across syllable positions. This paper reports the results of four experiments designed to study the perceptual similarity of mirror image acoustic patterns resembling the formant transitions and steady-states of the CV syllables /ba/, /da/, /ab/ and /ad/. Using a perceptual learning task, we found that subjects can learn to assign arrangements of these mirror image acoustic patterns more consistently to arbitrary response categories than similar arrangements of the patterns based on spectro-temporal commonalities. Subjects not only respond to the individual components or dimensions of these acoustic patterns but they also process the entire pattern and make use of its internal organization.

Some Experiments on Perceptual Learning of Mirror-Image Acoustic Patterns*

Mary Ellen Grunke and David B. Pisoni

In natural speech a single phonetic segment may have many different acoustic representations depending on the context in which it is spoken. An extensive body of research over the last thirty years has been directed at identifying acoustically invariant properties of phonemes which can mediate speech recognition. For example, stop consonants that follow a vowel are roughly the mirror image in time of their counterparts that precede the same vowel--the /b/ in the syllable /ba/ typically has rising formant transitions into the following vowel following release, whereas the /b/ in /ab/ has falling formant transitions. A similar, situation has been observed with /d/ in the syllables /da/ and /ad/--the /d/ is characterized by falling second and third formant transitions into the vowel and rising transitions out of the vowel.

A child who is acquiring language somehow learns to recognize consonants that occur in different positions as members of the same phonetic category. Can this acquisition process be mediated by acoustically-based similarity between mirror image patterns? That is, do mirror-image acoustic patterns share perceptually salient features that cause them to sound alike or to be classified together by listeners?

One approach to answering these questions has been to use speech sounds in a selective adaptation paradigm (e.g., Ades, 1974; Pisoni & Tash, 1975; Wolf, 1978). In general, this research has not found evidence for acoustic invariants as a basis for identification of stop consonants in different syllable positions. Ades (1974) reported that repeated presentation of a CV syllable had an adapting effect on a CV syllable continuum but not on a mirror-image VC continuum. The results of Pisoni and Tash's (1975) experiments on CV and VC syllable showed adaptation effects across position and suggested that there may be auditory property detectors that respond to rising or falling spectral information in the speech signal. In the absence of other information, rise-fall detectors could provide one way of categorizing the same consonant in pre- and postvocalic position. In a more recent study, Wolf (1978) considered a variety of acoustic properties including identical release bursts, mirror-image formant transitions, similar onset and offset spectra--and found none to be the basis for the invariant perception of place of articulation in initial and final syllable position.

In addition to use of the selective adaptation paradigm, another approach to the question of the perceptual relatedness of mirror-image acoustic patterns has been to use non-speech sounds in a similarity judgement task. Here again the data have not provided very convincing support for the hypothesis that mirror-image acoustic patterns are perceptually similar for a listener. Klatt & Shattuck (1975) and Shattuck & Klatt (1976) had listeners judge the similarity of brief pure-tone frequency glissandos. They found that similarity judgements of two-component patterns--either diverging, converging, both rising or both falling--were based primarily on the direction of the lower glissando component.

The approach adopted in the current experiments to the issue of mirror-image perceptual relatedness was to use nonspeech acoustic patterns of a slightly more complex nature than those used by Klatt and Shattuck. Our patterns contained a steady-state constant frequency (CF) interval in addition to a rapid frequency modulation. We also used a perceptual learning task in which listeners were trained to map four different acoustic patterns onto two response categories. The patterns were mapped according to the direction of the frequency transition, according to the position of the frequency transition relative to the steady-state, or according to a phonetic classification (i.e., mirror-image patterns paired with the same response). The question of principal interest was which mapping arrangement would produce fewest errors during acquisition. This learning task was used in Experiments 1 and 2 with a variety of sets of acoustic patterns in order to determine what perceptual features of the patterns were most salient to listeners in learning to group these patterns. In a third experiment listeners were asked to assign either phonetic or acoustic labels to the same acoustic patterns. This experiment provided a way of determining how accurately the patterns could be heard as speech or, alternatively, as nonspeech frequency modulated glissandos. Finally, in the last experiment similarity judgements were collected in order to directly compare our results with more complex signals with those obtained earlier by Klatt and Shattuck.

Experiment 1: Perceptual Learning Task

The first experiment used a perceptual learning task and compared acquisition performance among three response mapping conditions based on either a: (1) mirror-image relation, (2) the direction of glissandos, or (3) the relative temporal position of a glissando and steady-state frequency. Three sets of acoustic patterns, containing either single, double, or triple tones were used as test signals.

Stimuli

Three sets of stimuli containing four stimuli per set, were generated using a program that combines sine waves to form complex tones (Kewley-Port, 1976). The stimuli are shown schematically in Figure 1.

Insert Figure 1 about here

Each stimulus component consisted of a 60 msec linear rise or fall in frequency and a 140 msec constant-frequency portion. The four stimuli within a set differed in whether the frequency transitions were rising or falling and whether the transition preceded or followed the steady-state portion of the pattern.

The three stimulus sets also differed in the number of component tones, either one, two, or three. Frequency values were selected to correspond to values of the first, second, and third formants in the synthetic syllables /ba/, /da/, /ab/, and /ad/. For the single-tone set, the patterns were in the second formant region--1230 Hz. for the steady-state portion and transition endpoints of 995 for the "b" and 1465 Hz. for the "d" stimuli. For the double-tone set, the component frequencies corresponded to the second and third formants. The steady-state portion of the tone corresponding to a third formant was set at

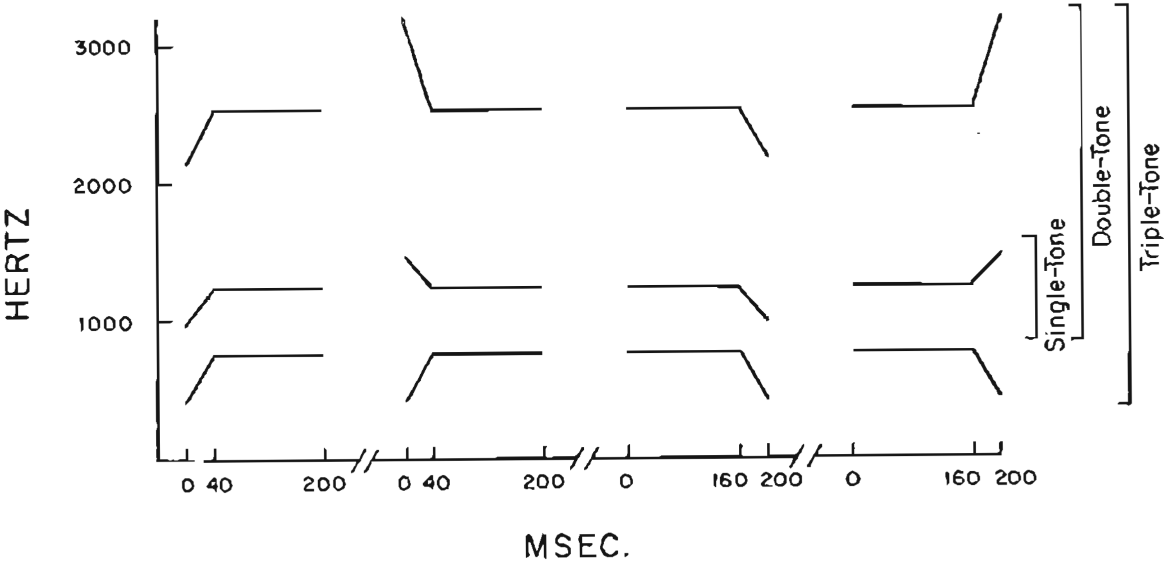


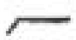
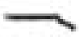
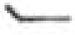
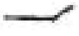
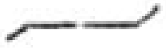
Figure 1. Schematic spectrographic patterns of the stimulus patterns used in Experiment I.

2525 Hz.; transition endpoints were 2180 Hz. and 3195 Hz., respectively, for the "b" and "d" stimuli. The lower tones of the double-tone stimuli were identical in frequency to the tones in the single-tone set. For the triple-tone set, all three formants were represented, although the frequency transition corresponding to the first formant always rose when it preceded the steady-state and fell when it followed the steady-state, in accordance with the formant motions observed in natural speech. The two upper components of the triple-tone stimuli were the same as the double-tone set; the lowest tone consisted of a 770 Hz. constant frequency portion and a frequency modulation to or from 410 Hz.

Experimental Procedure and Design

Stimulus presentation and data collection for all experiments reported here were controlled on-line in real-time by a PDP-11/05 computer. The auditory stimuli were presented to subjects at a comfortable listening level via TDH-39 headphones. Subjects were seated in front of a response panel in a sound attenuated cubicle. In Experiment 1 the response panel contained two response buttons, a feedback light above each response button, and a cue light that signaled the presentation of each tone 1 sec prior to its onset. The two response buttons were labeled simply "R1" and "R2".

The subjects task was to learn to map the four acoustic patterns in a given stimulus set onto two response categories according to a particular mapping arrangement. An experimental session consisted of 12 alternating study and test periods. During study periods, five repetitions of each stimulus were presented; first all "R1" repetitions were heard in random order, followed by all "R2" repetitions. The feedback lights indicated the correct response for each stimulus after it was presented. Test periods consisted of 20 trials, five per stimulus. On each trial, a stimulus was presented and subjects were required to respond by pressing one of the two buttons. The feedback lights then indicated the correct response. The order of the four stimuli was randomized throughout the test periods.

Subjects were assigned to one of three mapping conditions. In Condition 1, the mirror-image condition, stimuli with a rising transition preceding the steady-state () or a falling transition following the steady-state () were assigned to one response, whereas stimuli with a falling transition preceding the steady-state () or a rising transition following the steady-state () were assigned to the other response. In Condition 2, the rise-fall condition, the stimuli with rising transitions that either preceded or followed the steady-state () were assigned to one response, and the two stimuli with falling transitions were assigned to the other. Finally in

Condition 3, the temporal position condition, the stimuli were assigned to responses according to the temporal position of the transition, either preceding or following the steady-state portion. Subjects were never explicitly told the particular rule underlying the to-be-learned mapping arrangements although in some cases they were very obvious upon listening.

Nine separate conditions were formed by the factorial combination of three stimulus sets (single-, double-, and triple-tones) and three mapping arrangements (mirror-image, rise-fall, and temporal position). Ten subjects, recruited from introductory psychology courses at Indiana University, were randomly assigned to each condition. In this and all subsequent experiments, only subjects with no history of a speech or hearing disorder were used.

Results and Discussion

Of the three mapping conditions, temporal position showed the most accurate performance, with 88.4% correct pooled across all test trials. This mapping condition, of course, only required that listeners be able to distinguish the relative temporal order of a steady-state and frequency modulation. Of the two mapping conditions that required selective responding to particular properties of the frequency glissando, performance was better with mirror-image mapping (78.1% correct responses) than with rise-fall mapping (68.3%). With respect to the three stimulus sets, the triple-tone set produced slightly poorer performance (74.5%) than either the double-tone (81.6%) or single-tone sets (78.7%). Mean percent correct for each stimulus set and mapping condition are shown graphically in Figure 2.

Insert Figure 2 about here

An analysis of variance confirmed the statistical significance of the main effects of mapping condition, F(2,81) = 35.37, p < .001, and stimulus set, F(2,81) = 4.55, p < .05. Follow-up Scheffe tests of pairwise comparisons were performed using an overall significance level of .05. The post-hoc tests indicated that all three mapping conditions differed reliably from one another. However, with respect to stimulus set, only the double-tones were significantly better than triple tones in response accuracy.

The overall analysis of variance did not indicate a significant interaction between stimulus set and mapping arrangement, p > .10. However, the magnitude of the increase of mirror-image over rise-fall mapping was 16% and 22% for the single and double tones, respectively, and only 5% for the triple tones. In subsequent experiments with replication conditions we have observed very similar percentage increases in performance,

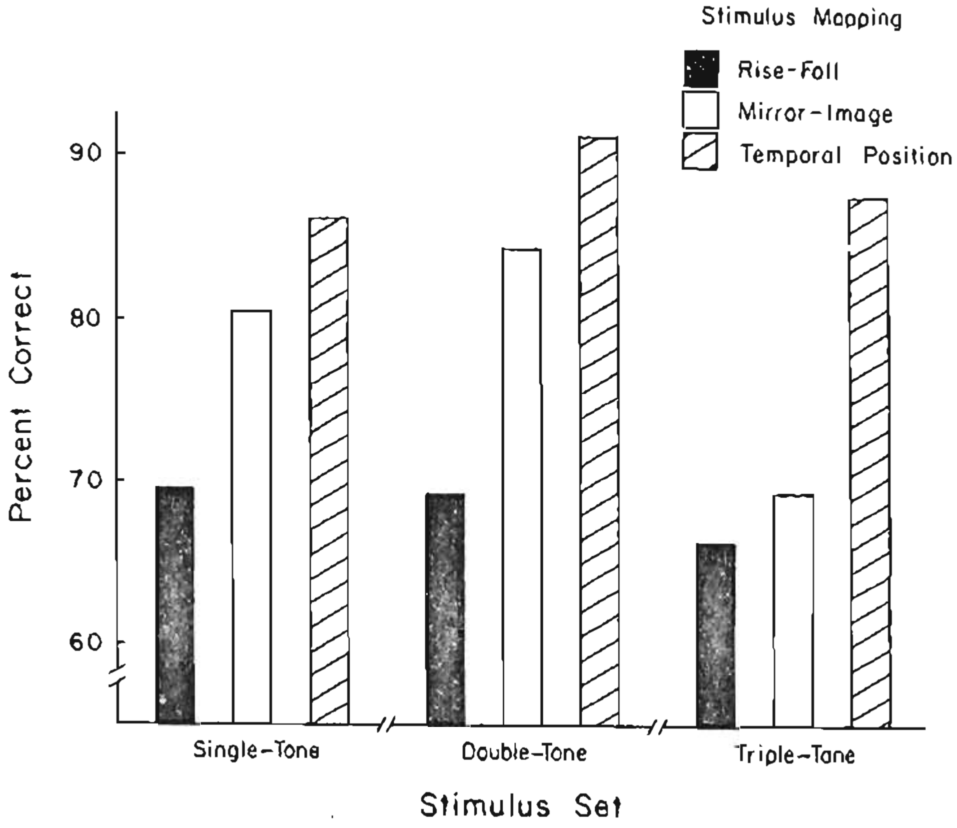


Figure 2. Percent correct performance pooled across subjects and individual learning trials for each stimulus set. The data are display separately for each response mapping condition.

i.e., relatively small improvement of mirror-image over rise-fall mapping with triple tones, and considerably larger improvement with the double tones.

In considering accuracy levels for the individual stimuli, performance was markedly poorer for transition-initial stimuli, 69.0%, than transition-final stimuli, 87.2% (F(1,81) = 167.02, p < .001). In Figure 3, percent correct performance is shown for the "ba," "da," "ab," and "ad" stimuli separately for each of the three mapping conditions.

Insert Figure 3 about here

Subjects in all three mapping conditions had little difficulty in assigning the final-falling ("ab") and final-rising ("ad") tones to the appropriate response category. Likewise, the higher performance level with temporal position mapping indicates that subjects could easily distinguish initial-transition stimuli from final-transition stimuli, a result that is not entirely suprising. The major source of difficulty in this task was learning to respond discriminatively to the initial-rising ("ba") and initial-falling ("da") signals. The superiority of mirror-image over rise-fall mapping demonstrates that it was easier to learn to map "ba" and "ab" tones to a common response and likewise "da" and "ad", than to learn that "ba" goes with

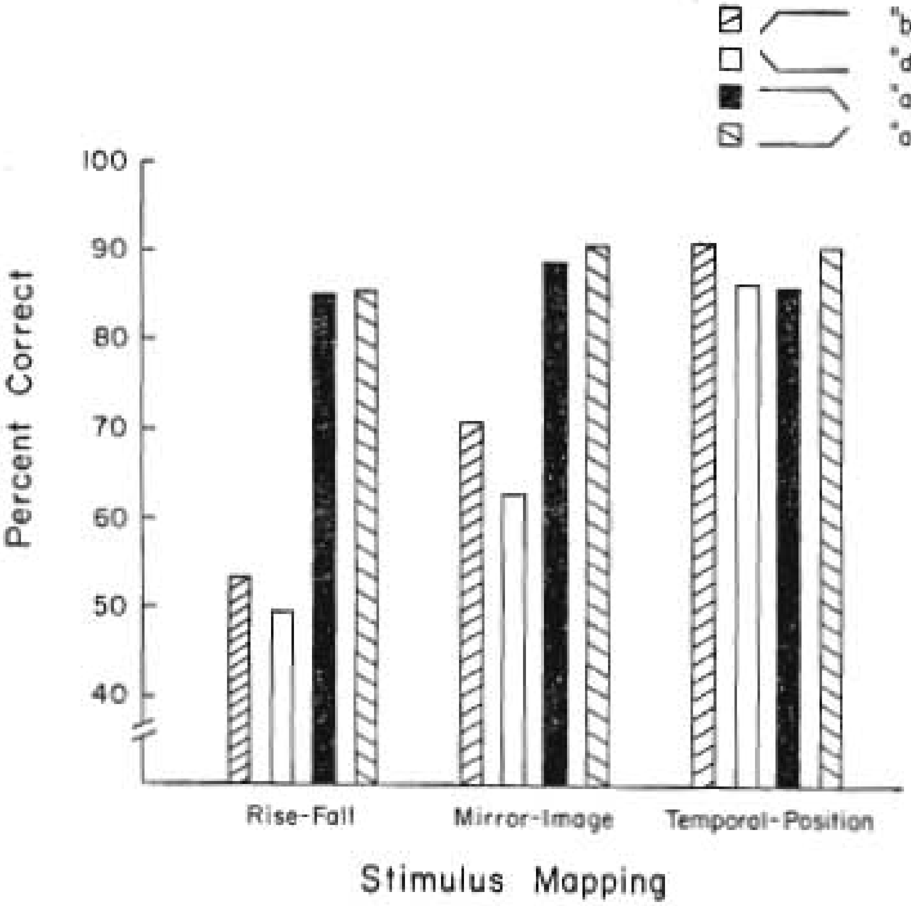


Figure 3. Percent correct performance for individual stimuli under the three mapping conditions of the experiment.

"ad" (both rising transitions) and "da" goes with "ab" (both falling transitions).

Experiment 2: Relevant Stimulus Properties

In view of the advantage of mirror-image mapping over rise-fall mapping, it was of some interest to determine what particular acoustic features or properties listeners were able to extract from the stimuli in the mirror-image condition. That is, what perceptually salient properties served as reliable discriminative cues for learning in the mirror-image condition? Two possible properties were investigated in Experiment 2. The first, short-term spectral composition of the frequency transitions, is specific to only the transitional portion of each signal. The second, frequency averaged across both steady-state (CF) and transition, is a configural property of each signal as a whole and may be equated roughly with the overall perceived pitch of the stimulus.

In the acoustic patterns considered thus far, both average frequency and transition endpoints were the same for signals within mirror-image pairs, but different across pairs. In contrast, transition endpoints and average frequency differed within rise or fall pairs, and thus could be considered an irrelevant cue for learning a rise-fall mapping response. In Experiment 2 two additional stimulus sets were used in which either transition endpoints or average frequency were held constant. Thus these properties would not be available as discriminative cues for any mapping arrangement. Would the advantage of the mirror-image over rise-fall mapping condition disappear under either of these new stimulus conditions? If elimination of a particular discriminative cue results in attenuation or reversal of the advantage of mirror-image over rise-fall mapping, then it can be argued that this cue contributes to, or underlies, the advantage of mirror-image mapping when the cue is present.

Method

The experimental procedure was basically the same as Experiment 1, but only the double-tone stimuli and two mapping arrangements, mirror-image and rise-fall, were used here. Subjects were trained with one of three stimulus sets which differed in whether steady-state frequencies, transition endpoints, or average overall frequency (pitch) were the same for all four stimuli in a given set.

The constant-steady-state set was the same as the double-tone set used in Experiment 1. Likewise the "ba" and "ab" stimuli were identical across the three stimulus sets. Only frequency values for the "da" and "ad" stimuli were modified in the remaining stimulus sets. In the constant-transition set, transition endpoints for each stimulus pattern were 995 and 1230 Hz for the lower tone, and 2180 and 2525 Hz for the higher tone. Thus, steady-state frequencies were 995 and 2180 Hz for the "da" and "ad" stimuli, and 1230 and 2525 Hz for the "ba" and "ab" stimuli. In the constant-average frequency set, the steady-state frequencies for the two "d" stimuli were 1183 and 2456 Hz for the lower and upper sinusoidal components, respectively; in the transition portion, the frequency increased or decreased to endpoints of 1418 and 2801 Hz.

Sixty additional undergraduate subjects who had not participated in Experiment 1 were again recruited from Indiana University's introductory psychology course. Ten subjects were randomly assigned to each experimental condition formed by factorally combining the two mapping arrangements and the three stimulus sets.

Results and Discussion

For both the constant-steady-state and constant-transition stimulus sets, the mirror image mapping condition again resulted in more accurate performance than the rise-fall mapping, 82.0% correct for mirror-image versus 68.0% for rise-fall with the constant-steady-state set, and 96.8% for mirror-image versus 64.9% for rise-fall with the constant transition set. However, when average frequency was held constant for all stimuli effectively reducing all salient cues, mirror-image mapping (68.3%) was not significantly different from, and was actually slightly poorer than, rise-fall mapping (73.2%). An analysis of variance on total correct responses substantiated the statistical significance of this interaction between stimulus set and mapping condition ($F(2,54) = 13.45$, $p < .001$).

With regard to performance on individual test stimuli, transition-final stimuli again showed higher response accuracy than transition-initial stimuli, but only for the constant-steady-state and constant-average-frequency stimulus sets. For the constant-steady-state set pooled over mapping conditions, response accuracy was 59.46% with transition-initial stimuli, compared with 90.50% for transition-final stimuli. For the constant-frequency set, response accuracy was 62.12% versus 79.38% for transition-initial and -final stimuli, respectively. When transition endpoints were the same for all four stimuli in the set (i.e., the constant-transition stimuli), response accuracy with transition-initial signals (82.46%) did not differ reliably from transition-final signals (79.25%).

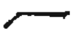
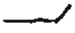
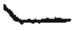
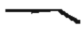
These results demonstrate the importance of the overall perceived pitch of the entire stimulus as a salient perceptual cue in learning to categorize acoustic patterns according to a mirror-image relation. Listeners did not respond simply to the individual components of these patterns, such as the transition or steady-state, but instead attended to a configural property of the entire stimulus--its overall pitch quality during perceptual learning.

Experiment 3: Labeling Task

The first two experiments used a perceptual learning task to study the acquisition of mirror-image acoustic patterns resembling speech. In Experiment 3 we used a labeling or identification task to examine another question concerning how the patterns are perceived: To what extent can these nonspeech signals, which were patterned after certain features of speech sounds, be heard as speech or alternatively, as frequency-varying tones? This question is relevant to the issue of whether subjects could deploy perceptual processes or strategies used in speech perception to solve the task.

Method

Subjects were required to identify acoustic patterns with either acoustic or phonetic labels. No feedback was provided in this experiment since we wanted to measure subjects' baseline ability to categorize these acoustic patterns in terms of the acoustic or phonetic attributes implicit in them without prior training or experience with these signals. In this experiment we were not interested in whether subjects could simply learn a sound-to-label association which may be arbitrary to them in the context of the test situation.

The stimulus patterns were the single-, double-, and triple-tone signals used in Experiment 1. Two conditions were examined. In the "phonetic" condition, subjects were told that the stimuli were distorted tokens of natural speech. The response labels were the syllables "ba," "da," "ab," and "ad," which were placed under four separate buttons on the response panel. In the "acoustic" condition, subjects were told that the stimuli were produced by a tone generator and consisted of a short interval with constant pitch, preceded or followed by a very rapid rise or fall in pitch. The response labels were schematic line drawings of the time course of the frequency change of each stimulus: (), (), (), ().

A factorial design was used again, with between-subject factors of stimulus set (single-, double-, or triple-tones) and label condition (acoustic or phonetic). Ten naive subjects were recruited from the same source as the previous experiments. They were assigned at random to the six experimental conditions. Sixty repetitions of each of the four stimuli were presented in random order to subjects, for a total of 240 labeling responses per subject.

Results and Discussion

Responses were scored as correct or incorrect depending on whether the indicated label was the most appropriate cue for the presented stimulus. Percent correct for both label conditions across the three stimulus sets are presented in Figure 4.

Insert Figure 4 about here

For the single- and double-tone stimuli, subjects were able to use acoustic labels more accurately than phonetic labels (single tones: 49.8% correct vs. 36.6% for acoustic and phonetic labels, respectively; double tones: 61.0 vs. 42.2%). However, with triple tones that contained energy in the first formant region, listeners assigned phonetic labels more accurately than acoustic labels (62.7% correct for phonetic labels compared with 42.6% for acoustic labels). That is, they were able to hear these patterns as speech. The interaction between label type and stimulus set was statistically reliable at the .01 level (F(2,54) = 6.78). Thus, listeners could attend selectively to either the phonetic or acoustic properties of these signals with better-than-chance accuracy, but their overall accuracy in doing so varied depending on the complexity of the signal and the specific properties attended to when under phonetic or acoustic instructions.

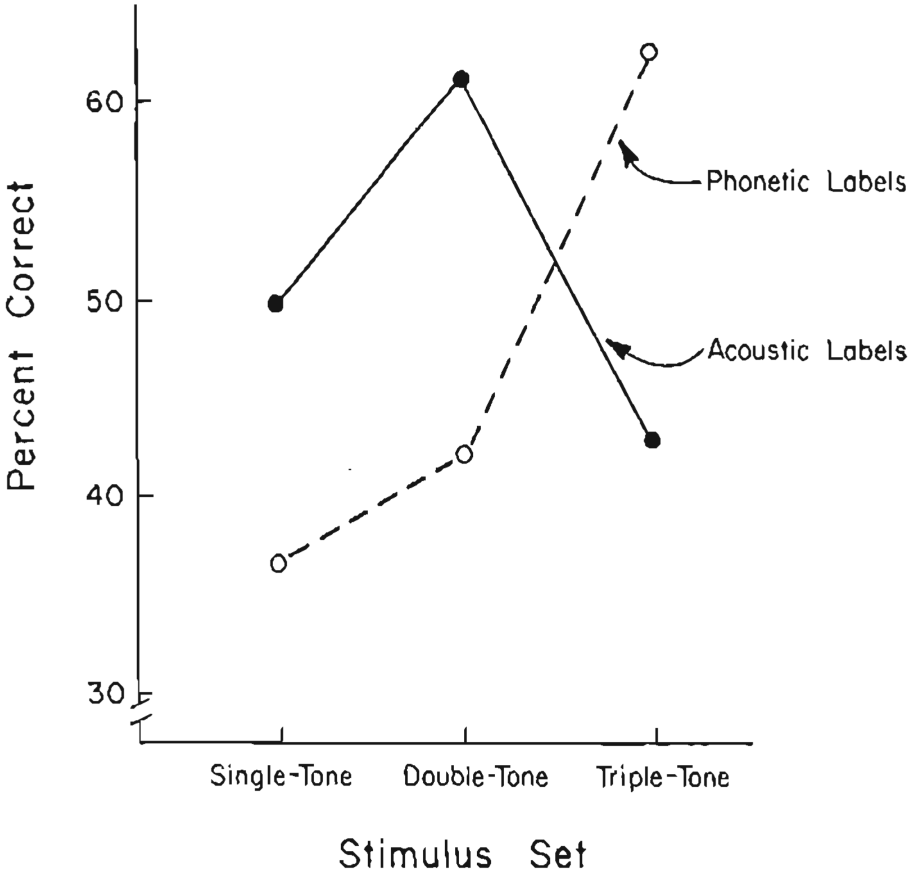


Figure 4. Percent correct identification in the labeling task for acoustic and phonetic labels. The data are shown separately for single-, double-, and triple-tone stimuli.

Examination of the labeling performance for the four separate stimuli indicated that for acoustic labels response accuracy was greater with transition-final signals ("ab": 69.44%, "ad": 62.83%) than transition-initial signals ("ba": 38.94%, "da": 33.39%), a familiar outcome observed in our previous learning experiments. Interestingly, however, when listeners attempted to assign phonetic labels to these signals, differences between transition-initial and transition-final signals were diminished substantially to a nonsignificant level; "ba": 45.17%, "da": 43.00%, "ab": 51.28%, "ad": 49.17%. These data are presented in Figure 5 where they have been pooled across single-, double-, and triple-tones because the observed pattern of responses was essentially the same for each stimulus set. Thus, these results demonstrate a dissociation between auditory and phonetic categorization of the same stimuli.

Insert Figure 5 about here

Experiment 4: Similarity Judgements

Klatt & Shattuck (1975) and Shattuck & Klatt (1976) found that two rising or two falling glissandos were judged to be more similar than mirror-image glissandos. Their stimuli differed from those used in the present studies, however, in that no

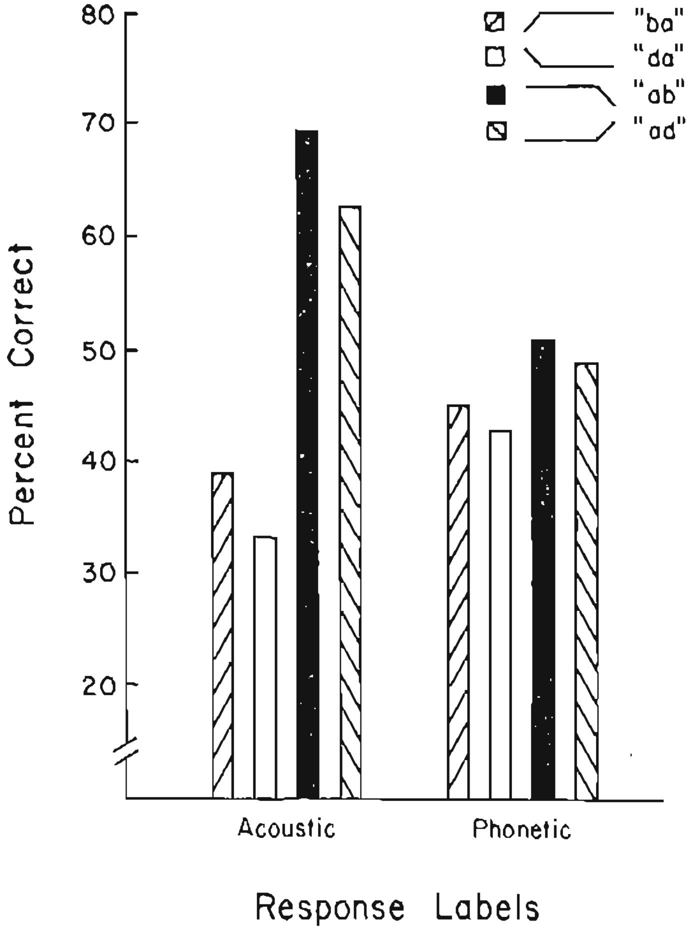


Figure 5. Percent correct identification for individual stimuli in the labeling task shown separately for acoustic and phonetic instructions.

steady-state (CF) portion was included. In Experiment 4 we repeated Klatt and Shattuck's similarity judgement task using our acoustic patterns. The question of interest was whether mirror-image signals would be perceived to be more similar, as results of the perceptual learning task suggest, or whether rise-fall pairs would be judged more similar, as found with glissando-only signals?

Method

On each trial subjects were presented three acoustic patterns separated by 250 msecs of silence. The subjects were instructed to indicate which stimulus sounded most different from the other two by pressing one of three buttons on the response panel. Each subject heard tones from both the constant-steady-state and constant-average-frequency sets used in Experiment 2. Trials with the two stimulus sets were blocked so that the first three blocks for a given subject contained tones from one stimulus set, and the next and final three blocks contained tones from the other stimulus set. The order of the stimulus sets was counterbalanced. Each of 24 possible stimulus triads was presented once in a block of trials.

In addition to experimental trials in which each of the three presented sounds were different (i.e., dissimilarity trials), we also included identity or catch trials in which two of the three stimuli were in fact the same. The identity trials were included in order to test for differences in the discriminability of the individual stimuli and also to check to be sure that subjects were in fact attempting to pick out the one stimulus from the triad that was most different from the other two. The odd stimulus in identity trials occurred equally often in each temporal position in a block of trials, and all thirty-six possible triad configurations for identity trials were used for each stimulus set in an experimental session. Thus, each block of test trials consisted of 24 dissimilarity trials randomly interspersed with 12 identity trials.

Eighteen additional subjects were recruited from a paid subject pool at Indiana University to participate in this experiment.

Results and Discussion

Only the results of the first half of each session, i.e., the initial three blocks of trials for each subject will be reported here. Subjects' responses on the second half suggested that experience with one stimulus set influenced similarity judgements on the other stimulus set. The task was difficult for subjects, as indicated by the finding that on 16% of the dissimilarity trials and 10% of the identity trials subjects failed to respond at all during the allotted 2.5 sec response interval. The results for the dissimilarity trials are therefore based only on the trials in which subjects gave a response. For the identity trials, however, percent correct was based on all trials unless otherwise noted.

Dissimilarity trials. The main question in this experiment was which stimulus pattern would be selected as most different from two other presented stimulus patterns--the stimulus which is not the mirror-image of either of the other tones, the stimulus which has a rising or falling transition in the direction opposite the transition of the other tones, or the stimulus which has a transition in a different temporal position, than the other tones? Several factors influenced subjects' judgements, including the particular stimuli presented on a trial, the order of the stimuli on a trial and, more importantly, the stimulus set used, whether it was the constant-steady-state or constant-average-frequency sets.

Figure 6 presents the percent responses for each response type (i.e., mirror-image, rise-fall, or temporal position) pooled over subjects and trials for each stimulus set.

Insert Figure 6 about here

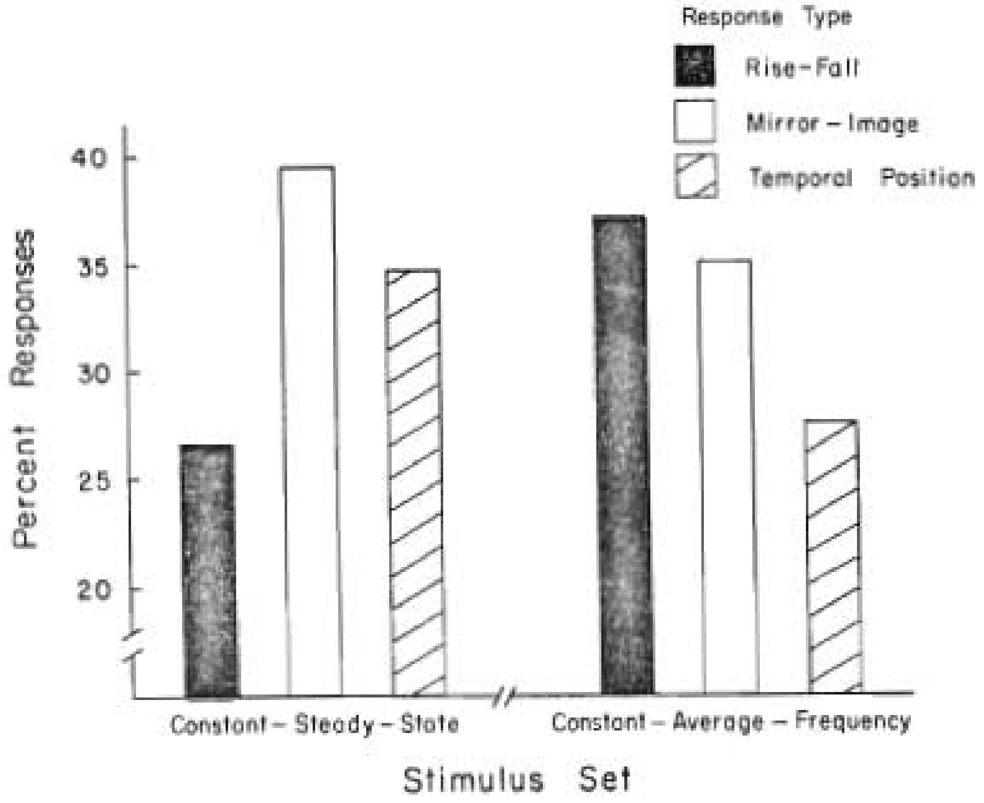


Figure 6. Distribution of responses obtained in the similarity judgement task across the two stimulus sets from Experiment 4.

With the constant-steady-state set, the largest proportion of responses, 39%, were in the direction of mirror-image similarity (i.e., the two nonselected tones were, in fact, mirror-images of one another). Thirty-five percent of the responses were based on the temporal position, and 27% on rise-fall. Individual subject performance, as contrasted with pooled data, indicated even more convincingly the predominance of mirror-image responses: Of the nine subjects run in this experimental condition, eight gave mirror-image responses most frequently, and only one gave temporal position responses most frequently.

With the constant-average-frequency set, in which overall average frequency was experimentally removed as a property that varied across stimuli, the pattern of results changed in the same manner as the results obtained in Experiment 2 involving a perceptual learning task. Rise-fall judgements were now slightly more prevalent (37%) than mirror-image responses (35%), which in turn were more prevalent than temporal position responses (28%). For the nine subjects who listened to the constant-average-frequency set, rise-fall was the predominant response of four of the subjects, mirror-image of four other subjects, and temporal position for only one subject.

Identity trials. Responses on identity trials were scored as correct or incorrect depending on whether the selected stimulus was in fact different from the other two presented on the trial. Overall percent correct on identity trials was 77.01%. Excluding omission responses, percent correct responses given that some response was made was 85.45%.

The identity trials were sorted into three categories according to whether the different, or "odd-out", stimulus was: (1) the mirror-image of the two identical stimuli, (2) had transitions in the same direction as the other two stimuli, or (3) had transitions in the same temporal position as the other stimuli in a triad. Based on responses from the dissimilarity trials, several predictions can be made about responses on identity trials. For example, with the constant-steady-state stimuli it would be expected that identity trials requiring a discrimination between mirror-image stimuli would be more difficult than trials requiring a discrimination based on a temporal position or rise-fall discrimination. This outcome is anticipated because mirror-image stimuli were judged as least dissimilar on the dissimilarity trials.

The observed pattern on identity trials for both stimulus sets was completely consistent with the data from dissimilarity trials: With the constant-steady-state stimuli, mirror-image trials showed poorer discrimination accuracy (76.85%) than either the rise-fall trials (82.41%) or temporal position trials (81.48%). With the constant-average-frequency stimuli, accuracy with rise-fall trials was slightly poorer (70.37%) than mirror-image trials (72.22%) which, in turn, was poorer than temporal position trials (78.70%).

To summarize the results from Experiment 4, responses on identity trials and dissimilarity trials converged to indicate the same conclusions: With pure-tone patterns that contain both a frequency transition and steady-state (CF), mirror-image pairs are judged to be more similar and are more difficult to discriminate than rise-fall pairs, provided that perceived pitch averaged over the entire pattern is available as a discriminative cue. When average frequency is selectively removed as a cue, rise-fall pairs tend to be judged as slightly more similar than mirror-image pairs, a result that is consistent with the findings from the previous perceptual learning experiments using the same stimuli. Thus the results from both experimental paradigms converge on the same conclusions.

Summary and Conclusions

Mirror-image acoustic patterns of the kind used in these experiments show an advantage in perceptual learning primarily because subjects respond not only to the individual components of these patterns but also to properties of the entire pattern in terms of its configural shape and internal organization. Subjects do not seem to attend selectively to only the gross shape of the spectrum at onset or offset but prefer instead to integrate and deploy salient acoustic cues contained in both the transitional and steady-state (CF) portion of the entire patterns in learning to assign the response categories consistently in this task. In the case of Mirror-Image patterns, the criterial differences between the response categories happen also to be correlated with salient and well-defined redundant properties of the patterns such as their overall perceived pitch which was an irrelevant and uncorrelated dimension of these same patterns when they were arranged in the Rise-Fall mapping condition.

It is also apparent from these results with nonspeech signals having properties simialr to those found in speech that differences in "mode of processing" can also control perceptual selectivity and influence the perception of individual components of the stimulus pattern as well as the entire pattern itself. This can occur in quite different ways with the same stimulus depending primarily on whether the subject's attention is directed to coding the auditory properties of the signals or the phonetic qualities of the patterns. In the former case, the process is more analytic involving the processing or "hearing out" of the individual components of the stimulus whereas in the latter the process is more nearly wholistic since the individual components can be combined to form a well-defined and highly familiar phonological category.

In summary, the overall results of our experiments indicate that acoustic patterns similar to pre- and post-vocalic variants of a particular stop consonant have more salient correlated properties in common with each other than other acoustic patterns that may have transitional movements in the same direction. These new results on mirror-image patterns have been obtained in a perceptual learning task despite the report that the perceptual similarity of these acoustic patterns cannot be recognized consciously by subjects as shown in earlier reports that measured the perceptual similarity of mirror-image acoustic patterns more directly. Further research on this problem is currently planned with infants, young children and monkeys to determine developmental trends and delimit cross-species differences in processing complex patterns resembling speech sounds.

References

Ades, A. E. How phonetic is selective adaptation? Experiments on syllable position and vowel environment. Perception & Psychophysics, 1974, 16, 61-66.

Kewley-Port, D. A complex-tone generating program. Research on Speech Perception Progress Report No. 3, 1976, Department of Psychology, Indiana University, Bloomington, Indiana, Pp. 203-208.

Klatt, D. H. and Shattuck, S. R. Perception of brief stimuli that resemble rapid formant transitions. In. G. Fant and M. A. A. Tatham (Eds.), Auditory Analysis and Perception of Speech. New York: Academic Press, 1975, Pp. 294-301.

Pisoni, D. B. and Tash, J. Auditory property detectors and processing place features in stop consonants. Perception & Psychophysics, 1975, 18, 401-408.

Shattuck, S. R. and Klatt, D. H. The perceptual similarity of mirror-image acoustic patterns in speech. Perception & Psychophysics, 1976, 20, 470-474.

Wolf, C. G. Perceptual invariance for stop consonants in different positions. Perception & Psychophysics, 1978, 24, 315-326.

Footnotes

*This research was supported by NIMH research grant MH-24027, NIH research grant NS-12179, an NIMH Post-doctoral fellowship to Mary Ellen Grunke and a fellowship from the Guggenheim Foundation to David B. Pisoni. An earlier report of some of these findings was presented at the Acoustical Society of America meetings in Cambridge, Mass., June 1979 by Mary Ellen Grunke.

ON SET-INDUCTION IN SENTENCE PERCEPTION

Robert E. Remez

Department of Psychology, Indiana University

Birgit Herbeck

Department of Linguistics, Indiana University

and

David R. Williams

Department of Linguistics, Indiana University

Abstract

A set-induction test was used to determine the relative perceptual salience of syntax, lexical structure, and accent pattern in auditorily presented sentences. Sentences were syntactically unambiguous, were of moderate syntactic complexity, and were otherwise normal but for the complete elimination of prosodic variation. The results showed a definite perceptual set for syntax and lexical boundary occurrence, but not for underlying accent pattern, and therefore disagree with previous research. The finding points to the importance of syntactic and lexical attributes in the perception of prosodically reduced sentences, and urges the use of ordinary sentences in experiments which study sentence perception.

In a recent review of psycholinguistic research, Johnson-Laird (1974) described the central problem of the enterprise as explicating the process of understanding sentences. Within this program, research has initially been concerned with grammatical variables, on the straightforward assumption that perceiving a sentence requires not only the recognition of the words but the assignment of logical relations among them as well. Experiments designed to reveal the nature of grammatical organization in sentence perception have typically taken the precaution of presenting the sentential "stimuli" in auditory form with subdued intonation, although in some cases visual presentation has been used. These methods of stimulus display prevent elements of prosody which might otherwise act as cues to the underlying syntactic organization (e.g., the modulation of vocal rate, pitch, amplitude, or rhythm occurring at constituent boundaries) from affecting the perceptual process. When a subject demonstrated syntactically organized percepts in the absence of explicit (prosodic) markers in the stimulus, this performance was interpreted as reflecting "psychologically real" syntactic mental operations (Fodor & Bever, 1965). As originally formulated, sentence perception was held to be a special grammatically-keyed activity which did not rely on crude cueing or prompting in the stimulus for its effectiveness.

The emphasis on the psychological imposition of logical organization on sentence elements has been retained in the more variegated conceptions of sentence perception advanced recently (e.g., Bever, 1970; Kimball, 1975; Marslen-Wilson & Welch, 1978; Wanner & Maratsos, 1978). Nevertheless, the continuing theoretical neglect of prosodic factors in the study of sentence perception has led at least one researcher to question the wisdom of attaching such preeminent importance to cognitive syntactic processes before the perceptual contributions of nonsyntactic sentence parameters have received appropriate investigation (Dooling, 1974). This point of view is grounded in the equally straightforward assumption that ordinary (as opposed to laboratory) sentence perception is directed toward utterances with fully expressed prosodic aspects; if prosody is perceptually prominent, a premise with strong intuitive appeal, then the formulation of the sentence perception process on the one hand ought to describe the coalescence of syntactic pattern and sentence "melody," and on the other hand should delineate the realtime processing constraints on sentence perception. Both of these issues have received research attention lately (see Haggard, 1975). The present investigation, based on the earlier study of Dooling (1974), is concerned with determining the relative perceptual salience of syntactic, lexical, and prosodic factors in sentence perception. Our intention here is to critically examine some of the experimental evidence that has drawn attention to prosodic dimensions of sentence perception.

The present study was motivated by the finding of Dooling (1974) that subjects prefer overwhelmingly to attend to sentence accent pattern rather than to sentence syntax; the test which reveals this is of the set-induction type thought to evoke syntactic processes in listeners (Mehler & Carey, 1967). Subjects in both the Mehler and the Dooling studies transcribed sentences masked by noise. In each case a perceptual set was established by presenting the subjects a number of structurally consistent sentences; this group of sentences may be said to comprise the _acquisition series_, for at its conclusion the listener is presumed to have established a tacit expectation that the structural consistency is characteristic of all sentences in the block. The last sentence of each block, called the _probe_, served to measure the strength of the induced expectation either by maintaining the structural consistency or by departing from it in precise ways. In control conditions, all sentences were of a single kind, both those of the acquisition series and the probe. But, in test conditions the probe differed from the acquisition series along a dimension of potential perceptual significance. The ability of listeners to correctly transcribe the probe when it differed from the acquisition set was assumed to vary with the perceptual importance of the differing dimension. When the probe sentence differed from the acquisition series on a perceptually crucial dimension, the acquisition series should have misled the listener in a perceptually costly way, and the transcription should be poorer than in control conditions. However, when the consistency was violated in a perceptually insignificant way, then the probe sentence should have been transcribed no more poorly than the control probes which do not differ from the acquisition sentences. The dependent measure was the accuracy of probe transcription, by sentence (Mehler) or by word (Dooling).

In the study which first established the use of this test, Mehler and Carey (1967) were interested in the relative perceptual salience of surface versus deep structural properties of sentences. In their test of surface structure salience, acquisition and probe sentences differed as do sentences (1) and (2):

(1) They are forecasting cyclones.

(2) They are conflicting desires.

In sentence (1), "cyclones" stands alone syntactically, while the two words "are forecasting" comprise the verb. In type (2), however, "conflicting desires" forms a single constituent, while the verb is the word "are." Different surface structure descriptions are assigned to these two sentences, and subjects who received type (1) as acquisition sentences and type (2) as the probe, or vice versa, showed a decrement in the accuracy of their transcriptions relative to performance in control conditions. When led to induce a mental set, the authors claimed, the subjects complied syntactically, demonstrating the perceptual importance of mental operations directed toward surface structure. However, the magnitude of the experimental effect in the deep structure condition was smaller, and problems were encountered in the control condition in that part of the experiment as well. Although the test was equivocal with respect to deep structure, it did nevertheless reveal that subjects became entrained to the surface structure of sentences.[1]

We may note, in addition, that the sentences used by Mehler and Carey were read at a normal rate but with monotone intonation. The reader of these sentences was well practiced in producing this kind of precautionary drone, "appropriate to no sentence in English." Further, we may observe that the scoring procedure was coarse grained, at the level of the entire sentence transcribed correctly.

In reply to the report of Mehler and Carey, Dooling (1974) questioned the importance of syntax in sentence perception studies. The arguments of Martin (1972) on the rhythmic integrity of connected speech led Dooling to suspect that prosodic factors might be of great perceptual importance, perhaps exceeding that of the processing strategies which purportedly assign deep and surface structure representations among sentence elements. To test this possibility, Dooling sought to determine the degree of disruption produced by differences in accent pattern between acquisition and probe sentences, relative to the effect of a syntactic difference in this paradigm. His test compared sentence types of the accent structure portrayed in Fig. 1.

****

Insert Figure 1 about here

****

Both sentences in the upper portion of the figure have the same syntax, but their accent patterns differ. The words "happy" and "people" are trochaic, while "precise" and "accounts" are iambic. A test using these two sentence types was designed by Dooling to reveal the importance or lack of importance of this nonsyntactic variable during sentence perception. But, to put the question of syntax versus prosody directly, sentences of the types in the lower part of Fig. 1 were also contrasted using the set-induction procedure. These two sentences differ in accent pattern and in syntactic structure. To force listeners to treat accent pattern as an abstract property of sentences, a monotone, stress-free reading style was used; the standard precaution for determining the psychological reality of grammatical processes was thus employed to detect the perceptual contribution of "abstract" prosody. Dooling argued that a comparison of the accent condition, accent plus syntax, and control conditions would then specify the perceptual relation of syntactic and nonsyntactic factors in sentence perception.

A reasonable assumption about the effect of simultaneous disruption of a perceptual set for accent and syntax is that the combined effect should exceed the effect of disruption of either accent or syntax alone. Accent pattern could be supposed to play a role in sentence perception, and the set-induction test could likewise be proposed as a reasonable indicator of this. Disruption of the set for syntax in addition to the set for accent pattern might then further degrade transcription performance. But, although subjects did produce poorer transcriptions when the accent set was violated than in control conditions (which at the very least establishes the contribution of accent pattern to sentence perception), performance was not any poorer when the sets for accent and syntax were disrupted together. Although the combined disruption of accent and syntax caused the mean of correctly transcribed words to decrease relative to the accent alone condition, this difference was not statistically significant. Both the accent condition and the accent plus syntax condition did differ significantly from the control condition in which acquisition and probe were of the same sentence types. In essence, then, Dooling had found that when accent pattern sets were disrupted the effect was so great that it mattered very little whether syntax was also disrupted. Put to this test, then, the Mehler and Carey (1967) finding seems questionable, though not on grounds of plausibility. Rather, the fact that rhythm was uncontrolled in that study raises the possibility that grammatical processes may have been overshadowed by more important perceptual processes keyed to accent pattern. The evaluation of these discrepant findings is the focus of our present study.

In our view, the findings of Mehler and Carey (1967) and of Dooling (1974) deserve replication with attention paid to technical problems in the original procedures which prevent clearcut interpretation. The present study seeks first to combine the two prior experiments in a single test of syntactic and nonsyntactic parameters, and second, to incorporate experimental precautions of a somewhat different sort to supplement those of earlier efforts. These precautions include a change in the dependent measure as well as a reform of the syntactic and the accent structures of the sentences. First, why should the role of syntactic processes have been so different in the two studies when each used sentences of the same apparent form, "They are ______________."? If we assume that sentences of this type do not provide the listener with a suitably rich syntactic object in either study, then the Mehler sentences may have actually exaggerated the role of syntactic processes, while the Dooling sentences may have attenuated it. Consider, from this perspective, that Mehler`s sentence types (1) and (2) above are implicitly puzzling, and seem to demand deliberate syntactic resolution, while Dooling`s sentences in Fig. 1 appear syntactically unambiguous. Although the sentences are superficially comparable in both studies, those used in Mehler`s study may have evoked an increased reliance on syntactic attributes in face of ambiguity; hence the greater performance decrement than found by Dooling when that same dimension of the set was disrupted between acquisition and probe. In addition, the sentences used in the accent study may have permitted subjects to adopt a pragmatic simplification strategy in place of syntactic operations. If such a strategy had identified items of the "happy people" type as _things_ and those of the "terrible to hear" type as _attributes_ but not as things in themselves, then the syntactic problem could have been solved nonsyntactically in a fashion fairly appropriate to the presentation----they are sneaky foxes; they are ugly muscles; they are pretty flowers; they are bashful children. This manner of display might encourage listeners to treat the predicates as list items rather than as proper sentence constituents.

Our experimental approach to this syntactic dilemma was to design sentences of greater length and clearer structure, with every attempt made as well to avoid semantic peculiarities. The new, improved sentences are as dissimilar as is possible within the constraints of syntactic conservation required by the set-induction test. In taking this precaution, we expected to prevent subjects from dozing syntactically; yet due to the unambiguous quality of the sentences, we hoped not to encourage any extraordinary grammatical efforts.

The test performed by Dooling was specifically intended to clarify the interaction of syntax and prosody in sentence perception. However, an examination of the kind of utterances used in this earlier study reveals that the test may not have been entirely appropriate for evaluating the role of these two factors. Despite Dooling`s expression of interest in rhythm as a source of nonsyntactic organization, the sentences in his experiment were read in a monotone, in a _syllable_-timed fashion at the rate of two syllables per second. This kind of utterance would not belong to the genre of _stress_-timed phenomena which Martin (1972; 1975) intended to describe. Martin attempted to formalize a set of correspondences among emphatic stress, syllabic accent, and the relative timing of the expression of these linguistic entities in the actual utterance. In his exposition, the level of stress assigned to any particular syllable is determined by its place in an abstract hierarchical description of the entire sentence; an underlying meter is assumed to regulate the equal intervals between strong stress beats at the sentence level. But in the Dooling experiment, accent was an entirely lexical affair in which the temporal parameters of syllabic expression were established by the metronome, not by a production hierarchy; the sound pattern simply did not include stress distinctions. This style of production was intended to insure the comparability of Dooling`s stimuli with Mehler`s. However, it also made the identification of accent pattern contingent on the identification of the words, since accent was not an explicit aspect of the stimulus. Accent pattern could be derived lexically, but it could not be the object of attention perceptually. Because the style of presentation was unaccented and temporally regular in the extreme, no explanation for the finding can be sought in recent studies of connected speech which might otherwise present evidence bearing on Martin`s proposal. Neither the experiments demonstrating a processing advantage for predictably stressed items (Shields, McHugh & Martin, 1974; Cutler, 1975; Cutler & Foss, 1977); nor those suggesting that timing parameters are a function of lexical or syntactic structure (Huggins, 1975; Klatt & Cooper, 1975; Hamill, 1976; Geers, 1978; Nakatani & Schaffer, 1978); nor those proposing that phonetic identifiability relies on the perceptual derivation of a metrical frame normalized for rate of production (Summerfield, 1975; Dorman, Raphael & Liberman, 1976; Port, 1977; Verbrugge & Shankweiler, 1977) can be thought to explain or support Dooling`s accent pattern effect. Each of those studies examines a dynamic property of speech perception, and abstract accent is properly a lexical matter and nothing more dynamic than that. Independent of this evidence, we do know that lexical accent can be a prominent dimension in linguistic tasks (Brown & McNeill, 1966; Robinson, 1977). This aspect of the sentences may have proven the most salient perceptual property by default in Dooling`s situation of reduced syntactic and semantic complexity. If so, the role played by accent pattern may not be as great in the perception of our more syntactically natural sentences due to the variety of information at more abstract levels. That perception may be geared naturally to the "figural" properties of stimulation rather than the "elemental" units is an argument persuasively presented most recently by Runeson (1977).

One other precaution was taken, in this case against potential floor effects. This seemed prudent because the average performance in Dooling`s most difficult condition was .41 keywords (of a possible 2.0) transcribed correctly. This means that many subjects must have completely failed in this condition, rather than merely performed more poorly due to the experimental manipulation, as the prediction declared. Given the natural shyness of the average subject in psychology experiments and the intrinsic difficulty of perceiving sentences masked by noise, a more sensitive procedure might, first, widen the range of potential performance by increasing the sentence length, then encourage guessing on the part of the listener, and, finally, proceed syllable-by-syllable in scoring. These steps increase the distance between ceiling and floor, and provide a finer-grained method of tracking subject performance.

Our new sentences, of eleven syllables each, are controlled for accent pattern in a manner analogous to Dooling`s materials; these longer sentences will provide a truer test of the interaction of accent and syntax because of the increased syntactic complexity and the more varied accent pattens offered to the listeners. However, it will not be appropriate to include explicit prosody as a sentence parameter until the original effect noted by Dooling--the salience of abstract accent pattern--is established in the context of syntactic complexity. Therefore, the sentences in the present test are read metronomically, with flat intonation and no stress contrasts.

A third type of set-induction is also tested here, on the assumption that subjects in this kind of noise-masked transcription task may also be attending to the regular occurrence of lexical boundaries at particular points in the syllable sequence. This factor had been considered by Dooling as a control condition, and he reported no set effect in that case. Because his test involved the assignment of a single lexical boundary between keyword syllables 1 and 2, or between 2 and 3, though, the task was quite proscribed. As such it may not have been a fair test of this intermediate-grained factor, which is coarser than syllabic accent yet finer than syntactic constituent. The inclusion of this third factor was felt to provide the opportunity for an independent check on the validity of the test, as well as to offer an indicator of the level of abstractness of the subject`s attention to the dimensions of sentences in the set-induction test. In summary, our test uses longer sentences, of unambiguous syntax and varied accent pattern, to determine the relative perceptual salience of syntactic constituents, abstract accent pattern, and lexical boundaries occurring in the syllable series.

Method

Materials. Four groups of sentences of eleven syllables each were designed to satisfy the conditions that syntactic structure, accent pattern, and occurrence of lexical boundaries be consistent within each group. Types I and II were used to test the magnitude of syntactic set-induction, for, as can be seen in Figs. 2 and 3, these two types differ in syntactic description. The contrast between these two types is simple, involving neither a change in abstract accent pattern nor in the particular syllables between which lexical boundaries occur. Figs. 4 and 5 show the accent patterns and lexical boundary locations for the four sentence types.

****

Insert Figure 2 about here

****

****

Insert Figure 3 about here

****

****

Insert Figure 4 about here

****

****

Insert Figure 5 about here

****

A test of set-induction of accent pattern is constructed with sentence types I and III, as Fig. 4 shows; each key word in type I differs in accent pattern from the corresponding word in type III. The contrast is simple because the two types of sentences share the same syntax and have coincident word boundaries in the series of eleven syllables.

Set-induction of the occurrence of lexical boundaries in the syllable series is tested by using sentence types II and IV, as can be seen in Fig. 5. These two sentence types have the same syntactic description as well as the same pattern of strong and weak accents. Only the location of word boundaries differs between them, permitting this simple contrast.

Each of the four sentence types was used in a separate control condition to establish baseline performance for set-induction. Control conditions used the same sentence type for the acquisition series as for the probe sentence. Eleven sentences were required to make the control test in which each sentence type was used: ten to make the acquisition series, and one to make the probe. Tests of the effect of breaking the set also required eleven sentences of each participating type. In one instance the acquisition series of ten was of one type and the single probe of the other; this arrangement was reversed in the counterbalanced instance. Types I and II were used in two conditions other than the control, and III and IV were each used in a single noncontrol test, making the total sentence inventory 33 of each of the first two types, and 22 of each of the second two. In addition, a set of fifteen eleven-syllable practice sentences of uncorrelated syntax, accent pattern and lexical boundary occurrence was made. This set was used to introduce the task and to provide a check on the intelligibility of the freely varying sentences against the preset standard of 60% of the syllables correctly transcribed. (This standard was established during piloting of the study as a suitable criterion for avoiding ceiling and floor effects.)

Each sentence used in the experiment was typed on one line of a three-by-five card, and the deck was shuffled to prevent introduction of bias in the reader. The reader, who sat in a sound-insulated booth, read each sentence in turn, covertly rehearsing on many occasions, and then speaking each sentence in a monotone; every syllable received subjectively equal emphasis. For each sentence, the production of each syllable was synchronized to the isochronous flashes of a Franz "Flash-Beat" metronome (LMFB4) set at 120 (one flash each half-second) and placed in view outside the booth with the sound maker defeated. An Electrovoice D054 microphone and an AMPEX AG-500 tape recorder set at 7.5 ips were used to record these utterances onto the master tape. The reader, an adult male native speaker of American English, was naive to the purpose of his task.[2]

Stimuli. A "script" for the individual conditions was created for the fifteen sentences of the pretest, the four control conditions, and the six set-disruption conditions. Sentences of the appropriate types in each condition were randomly ordered, and the individual tests were assembled by remastering the original recording to match the script. The ordering of conditions was not random, however, but followed the principle that contiguous tests should not use sentences of the same type. The script appears in the Appendix.

Channel A of the test tape was used for the speech track. On Channel B a noise mask was added, aligned temporally with each sentence, by recording the output of a Grason-Stadler Noise Generator (1725) set at the 10KHz bandwidth position. The noise was onset approximately 1.5 seconds before the speech in each trial and was offset approximately 1 second after the end of the last syllable in the sentence. Before each session the playback channels were set individually to 68dB SPL for noise and 64dB SPL for speech, and then mixed to provide a binaural signal to subjects, who listened over Grason-Stadler TDH39-500 headsets. (The particular attenuation settings for each channel had been established during piloting of this study using the practice block of sentences; mean transcription accuracy was thereby adjusted to 60% of the syllables correctly identified on the fifteenth sentence.)

Subjects. Sixteen listeners were drawn from the Indiana University Psychology Department Subject Pool. All were right-handed native speakers of English with no history of impaired speech or hearing, and each subject further stipulated that he or she had scored 550 or better on the verbal portion of the Scholastic Aptitude Test. This last criterion was imposed to insure that the subjects were reasonably adept in using English. All subjects received course credit in exchange for their participation.

Procedure. Testing was done in four groups. Each subject sat at a carell in a sound-attenuating room and transcribed sentences into prepared booklets. Subjects changed pages after each sentence, and changed booklets at each new block of trials.

Subjects were instructed that sentences would be played over the headsets, but that the message of each might be difficult to hear because the recording was deliberately noisy. They were told to write down the sounds that they heard even if they could only identify part of a sentence. The correct transcription of the first practice sentence was divulged, to orient the subjects to the task, and at that point the practice test was begun.

After the pretest, the subjects were questioned about any difficulties encountered, and when it was ascertained that each subject could indeed perform the task, the first five blocks of trials were run. A short intermission, during which subjects did not discuss the test, intervened between the first five and the last five blocks. Upon conclusion of the test, subjects were asked to reflect about the test sentences. When the design of the test materials was explained by the experimenter, no subject indicated that he or she had detected the underlying differences between the sentences.

Results and Discussion

Performance in control and set-disruption conditions was scored as number of syllables correctly transcribed of each probe sentence. Intrusions were not considered to be errors, but order inversions were. To test the consistency of performance within each condition a Cronbach reliability analysis was performed. For the three test conditions and the control conditions the correlations were positive [Cronbach alpha = .499 (CONTROL), alpha = .167 (SYNTAX), alpha = .382 (ACCENT PATTERN), alpha = .604 (LEXICAL BOUNDARIES)]. On the basis of this consistency of performance across subjects and conditions, four scores were then derived for each subject: one control score, the average of the four instances of that test; and three set-disruption test scores, one from each of those three conditions, each score the average of the two trials in each test. A one-way Analysis of Variance found transcription performance to differ significantly as a function of the test condition [F(3,45)=12.768, p<.001]. Dunnett`s tests comparing each set-disruption condition with the control revealed that performance was significantly poorer when syntactic structure or location of word boundaries in the syllable sequence had been disrupted (Dunnett`s d'=1.426; p<.01). Performance was not poorer, however, when accent pattern differed between acquisition and probe sentences. Table 1 lists the mean scores in each condition. In summary, we found that disrupting abstract accent pattern in the set-induction test does not cause performance to suffer, but that changing syntactic structure or location of word boundaries in the same circumstance does.

****

Insert Table 1 about here

****

It is not surprising that syntactic sentence properties figured prominently in our test. Because the sentences were deliberately read with no prosodic variation there was little else beyond the phonetic segments to which listeners could attend. That our syntactically and semantically enriched sentences elicited a different pattern of results than was found by Dooling (1974) is most likely the reflection of the change in salience of structure and meaning relative to production-related sentence properties such as accent and rhythm. In the earlier study by Dooling, sentences were impoverished in both dimensions, but in the present study, syntax and meaning approached a normal, ordinary level, while stress remained reduced to preserve methodological comparability with Dooling`s experiment. Of course, an utterance in which levels of stress are differentiated would probably provide a stronger stimulus for the perception of accent pattern and rhythmic figure. A test of this hypothesis, that prosodically explicit sentences evoke a rhythmic orienting response, will be quite simple to undertake by presenting the sentences used here in a prosodically normal form. Under such revised circumstances it may be possible to determine the minimal explicitness of rhythmic expression which can <u>attract</u> the perceiver`s attention. Evidence that accent or rhythmic set-induction is possible using ordinary, structurally unambiguous sentences with unconstrained prosodic expression would argue for a significant role of prosody in sentence perception. However, we must disagree with Dooling`s earlier conclusion that abstract accent pattern is either relevant to issues raised by Martin (1972; 1975) or that it is of particular perceptual salience in this kind of procedure. It may be quite difficult for subjects to develop a set for accent pattern under conditions of reduced prosody; it should be noted, though, that listeners are able to exaggerate the perceptual salience of an elemental aspect of meaningful material merely by being instructed to do so (Morris, Bransford & Franks, 1977). Whether or not abstract accent pattern or even explicit sentence rhythm can be isolated perceptually remains to be decided experimentally.

The significance of the disruption of a set for the coincident location of lexical boundaries across acquisition sentences further demonstrates that subjects appear to find abstract sentence properties to be of greatest perceptual value in this syllable-timed presentation. If subjects had listened syllable-by-syllable to each sentence and only then had solved the lexical recognition problems, we would have seen no such effect of disruption of the lexical boundary set; the assignment of word boundaries would have been a leisurely matter of secondary inference for the probe as well as for the acquisition sentences. But, because the acquisition sentences were structurally consistent, both in number of syllables in comparable words and in the location of word boundaries in the syllable series--and because listeners seemed to value this consistency--we may conclude that this pattern across sentences was actively used as a real-time constraint in word perception.

Lexical accent pattern might also be of similar value in this sort of perceptual circumstance, but perhaps it was simply not expressed strongly enough here to have been perceptually significant. Alternatively, it could be argued that in view of the overriding syntactic and semantic constraints, lexical accent may essentially be trivial in perception in most conditions other than those highly ambiguous ones which subjects encounter on occasion in the laboratory. Further research on these matters is required to confirm our speculation.

If, finally, as Lashley (1951) observed, rhythm is powerfully entraining----"one falls into step with a band, tends to breathe, and even to speak in time with the rhythm (p.127)"----then we may expect explicitly patterned sentences to elicit the effect which we did not find. Additional study of this problem may also show whether rhythm is a cue to syntax, or a rate-normalization aid to phonetic identification, or a paralinguistic variable oblique to the activities of sentence perception. Perhaps, rhythm is truly an ingredient and not a vehicle of communication. In conclusion, this study of syntax, word boundary placement and abstract accent pattern shows that typical prosodic precautions taken with otherwise ordinary sentences can lead subjects to favor abstract properties in their perceptual processes. In addition, the use of varied sentences of unambiguous but ordinary structure is advocated, for the type of attention subjects pay to sentence attributes may depend as much on the nature of those sentences as on the "psychologically real" syntactic operations of the perceptual process.

FOOTNOTES

Address correspondence to Robert E. Remez, Department of Psychology, Indiana University, Bloomington, Indiana 47405.

It is a pleasure to thank David B. Pisoni for extending to us the use of his laboratory to conduct this study. We are also indebted to Steve Simnick for his careful reading of our sentences, to William Badecker for his syntactic counsel, to Mickey Stentz for his statistical advice, and to Steve Braddon, Mary Smith, and Bill Twyford for their editorial aid.

[1] Sentence types in the deep structure tests resembled (i) and (ii):

(i) They are hesitant to travel.

(ii) They are difficult to employ.

These differ in the status of the pronoun "they." In (i) "they" is the logical subject (agent) of the imbedded verb "travel," while in (ii) it is the logical object (patient) of the imbedded verb "employ." This structural/logical difference in the sentences, however, does not involve a repartitioning of the strings into different constituents, as was the case in the surface structure pair. Rather, for Mehler and Carey, describing the differences between (i) and (ii) required resort to the deep structure level. On this account, it should be mentioned that the tools of structural analysis provided by syntacticians have undergone considerable refinement since Mehler`s experiment (Jackendoff, 1977), to the extent that the differences between (i) and (ii) would be apparent (c. 1979) in the <u>surface</u> structure descriptions. Perhaps a more correct restatement of their finding, giving Mehler and Carey full benefit, would be that syntactic processes which do not shift phrase boundaries relative to the lexical items can nevertheless induce a mental set. The significance of this proposal, though, awaits experimental test, due to the methodological problems with the "deep structure" trials mentioned already.

[2] It was not technically possible to control for precise syllable-by-syllable temporal variation except informally. However, had such opportunity been presented, it would have by no means been obvious which parameters of the acoustics to regulate nor precisely how they ought to have been arranged to construct perfectly stress-free isochronous utterances. This is due in part to the differences in intrinsic acoustic intensity (Lehiste, 1976) and intrinsic acoustic duration (Klatt, 1976) of phonetic segments, and also due to the problem of specifying the acoustic details of phenomenal isochrony in speech (Allen, 1972a, 1972b; Morton, Marcus & Frankish, 1976; Fowler, in press).

REFERENCES

Allen, G.D. The location of rhythmic stress beats in English: An experimental study, I. Language and Speech, 1972a, 15, 72-100.

Allen, G.D. The location of rhythmic stress beats in English: An experimental study, II. Language and Speech, 1972b, 15, 179-195.

Bever, T.G. The cognitive basis for linguistic structures. In J.R. Hayes (ed.), Cognition and the Development of Language. New York: Wiley, 1970. Pp. 279-362.

Brown, R., and McNeill, D. The "tip of the tongue" phenomenon. Journal of Verbal Learning and Verbal Behavior, 1966, 5, 325-337.

Cutler, A. Sentence stress and sentence comprehension. Unpublished Ph.D dissertation, University of Texas, 1975.

Cutler, A., and Foss, D.J. On the role of sentence stress in sentence processing. Language and Speech, 1977, 20, 1-10.

Dooling, D.J. Rhythm and syntax in sentence perception. Journal of Verbal Learning and Verbal Behavior, 1974, 13, 255-264.

Dorman, M.F., Raphael, L.J., and Liberman, A.M. Further observations on the role of silence as a cue for stop consonants. Journal of the Acoustical Society of America, 1976, 59, S40.

Fodor, J.A., and Bever, T. The psychological reality of linguistic segments. Journal of Verbal Learning and Verbal Behavior, 1965, 4, 414-420.

Fowler, C.A. "Perceptual centers" in speech production and perception. Perception & Psychophysics, in press.

Geers, A.E. Intonation contour and syntactic structure as predictors of apparent segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 273-283.

Haggard, M. Understanding sentence understanding. In A. Cohen and S.G. Nooteboom (Eds.), *Structure and Process in Speech Perception*. New York: Springer Verlag, 1975. Pp. 3-15.

Hamill, B.W. A linguistic correlate of sentential rhythmic patterns. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, *2*, 71-79.

Huggins, A.W.F. On isochrony and syntax. In G. Fant and M.A.A. Tatham (Eds.), *Auditory Analysis and Perception of Speech*. New York: Academic Press, 1975. Pp. 455-464.

Jackendoff, R. $\bar{X}$ syntax: a study of phrase structure. *Linguistic Inquiry Monograph* 2, 1977.

Johnson-Laird, P.N. Experimental psycholinguistics. *Annual Review of Psychology*, 1974, *25*, 135-160.

Kimball, J. Predictive analysis and over-the-top parsing. In J. Kimball (Ed.), *Syntax and Semantics*, *vol. 4*. New York: Academic Press, 1975. Pp. 155-179.

Klatt, D.H. Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 1976, *59*, 1208-1221.

Klatt, D.H., and Cooper, W.E. Perception of segmental duration in sentence context. In A. Cohen and S.G. Nooteboom (Eds.), *Structure and Process in Speech Perception*. New York: Springer Verlag, 1975. Pp. 69-89.

Lashley, K.S. The problem of serial order in behavior. In L.A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior: The Hixon Symposium*. New York: Wiley, 1951. Pp. 112-136.

Lehiste, I. Suprasegmental features of speech. In N.J. Lass (Ed.), Contemporary Issues in Experimental Phonetics. New York: Academic Press, 1976. Pp. 225-239.

Marslen-Wilson, W.D., and Welch, A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology, 1978, 10, 29-63.

Martin, J.G. Rhythmic (hierarchical) versus serial structure in speech and other behavior. Psychological Review, 1972, 79, 487-509.

Martin, J.G. Rhythmic expectancy in continuous speech perception. In A. Cohen and S.G. Nooteboom (Eds.), Structure and Process in Speech Perception. New York: Springer Verlag, 1975. Pp. 161-177.

Mehler, J., and Carey, P. Role of surface and base structure in the perception of sentences. Journal of Verbal Learning and Verbal Behavior, 1967, 6, 335-338.

Morris, C.D., Bransford, J.D., and Franks, J.J. Levels of processing versus transfer appropriate processing. Journal of Verbal Learning and Verbal Behavior, 1977, 16, 519-533.

Morton, J., Marcus, S., and Frankish, C. Perceptual centers (P-centers). Psychological Review, 1976, 83, 405-408.

Nakatani, L.H., and Schaffer, J.A. Hearing "words" without words: Prosodic cues for word perception. Journal of the Acoustical Society of America, 1978, 63, 234-245.

Port, R.F. The Influence of Speaking Tempo on the Duration of Stressed Vowel and Medial Stop in English Trochee Words. Bloomington, In.: Indiana University Linguistics Club, 1977.

Robinson, G.M. Rhythmic organization in speech processing.

Journal of Experimental Psychology: Human Perception and Performance, 1977, 3, 83-91.

Runeson, S. On the possibility of "smart" perceptual mechanisms. Scandinavian Journal of Psychology, 1977, 18, 172-179.

Shields, J.L., McHugh, A., and Martin, J.G. Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. Journal of Experimental Psychology, 1974, 102, 250-255.

Summerfield, A.Q. How a detailed account of segmental perception depends on prosody and vice-versa. In A. Cohen and S.G. Nooteboom (Eds.), Structure and Process in Speech Perception. New York: Springer Verlag, 1975. Pp. 51-68.

Verbrugge, R.R., and Shankweiler, D. Prosodic information for vowel identity. Journal of the Acoustical Society of America, 1977, 61, S39.

Wanner, E., and Maratsos, M. An ATN approach to comprehension. In M. Halle, J. Bresnan and G.A. Miller (Eds.), Linguistic Theory and Psychological Reality. Cambridge: MIT Press, 1978. Pp. 119-159.

FIGURE CAPTIONS

Figure 1. Sentences from Dooling (1974), used for testing accent set-disruption (top) and accent-plus-syntax set-disruption (bottom).

Figure 2. Syntactic descriptions of the surface structure of types I and III.

Figure 3. Syntactic descriptions of the surface structure of types II and IV.

Figure 4. Accent pattern and word boundary locations of types I and III.

Figure 5. Accent pattern and word boundary locations of types II and IV.

TABLE CAPTION

Summary of transcription performance.

TABLE 1

Mean Transcription Performance (Syllables Correct)

Condition	Mean
Control	9.078
Syntactic Set disrupted	6.688*
Word Boundary Set disrupted	6.875*
Accent Pattern Set disrupted	8.500

*p<.01; Dunnett`s d'=1.426

APPENDIX

This is the script of the test. The type of set disruption being tested is listed at the top of each block. The probe sentences are starred at the bottom of each.

PRETEST

The hamburger is a basic source of food.

Canada has closed her borders to felons.

The sweater had been attacked by hungry moths.

"Please don't play the trombone now," said Washington.

For relief of tension, Carter meditates.

The gypsies in Budapest wear bandanas.

No one could understand why Camille did it.

John brought a toaster and two kettles with him.

The lamb caught in the thicket could not escape.

Every autumn the chief orders pumpkin pie.

Thirty miles to the gallon is terrific.

Putting too much pepper on food can be bad.

Basketball is native to America.

The present King of France is bald and toothless.

*Several muddy rivers run through New Jersey.

SYNTAX TEST

Feeble arguments are greeted with laughter.

Foreign anarchists were captured in Boston.

Hardened criminals were questioned in meetings.

Prickly cactuses were planted in Texas.

Summer bungalows are rented at Easter.

Older Cadillacs were driven on hightest.

Tuna sandwiches were toasted in ovens.

Public finances were squandered on junkets.

Faithful royalists were banished from Togo.

Unripe canteloupes were purchased in Europe.

*Science verified that planets are distant.

CONTROL TEST FOR TYPE III

Enraged subversives were betrayed in Detroit.

Concealed revolvers are allowed in Quebec.

Absurd invectives are inscribed on balloons.

Enforced detention is prescribed for recruits.

Proposed improvements are required for sedans.

Minute contraptions are designed in Japan.

Concise restatements are desired on exams.

Unique inventions are displayed at bazaars.

Profound delusions were revealed through mistakes.

Prolonged engagements were secured at resorts.

*Mature opinions are expressed in debates.

WORD BOUNDARIES TEST

Pershing fantasized that conquest was simple.

Inquests document that torture is brutal.

William testified that Jennie was guilty.

Students realize that failure is painful.

Krogers advertized that lettuce is healthful.

Mothers fabricate that dentists are painless.

Leaders formulate that labor is costly.

Parents predicate that college is useful.

Moses prophesied that Canaan was lovely.

Theories explicate that oceans are shrinking.

*He concluded that sardines are delicious.

ACCENT PATTERN TEST

Local artichokes are eaten in autumn.

Porno magazines were outlawed in Cleveland.

Flower catalogues are printed on newsprint.

Changing attitudes are noticed toward music.

Costly merchandise was stolen from drugstores.

Heavy particles were filtered from mixtures.

Printed documents were copied with cameras.

English lemonade is sweetened with honey.

Ancient murderers were sentenced in public.

Comic messages were posted on Broadway.

*Reversed decisions were upheld in Iran.

CONTROL TEST FOR TYPE II

Mozart typifies that music is pretty.

Surveys indicate that children are careless.

Congress legislates that voters be sober.

Pastors intimate that friendship is sacred.

Brokers stipulate that prices are stable.

Illness symbolized that demons were present.

Shakespeare dramatized that virtue is tragic.

Experts certify that reading is easy.

Studies demonstrate that monkeys are clever.

Doctors specify that smoking is harmful.

*Rabbits recognize that carrots are tasty.

CONTROL TEST FOR TYPE I

Tender venison is roasted with charcoal.

Famous submarines were ruptured through error.

Frightened islanders were tortured with fire.

Science libraries were painted with latex.

Thrilling secrecies are mentioned at weddings.

Private property is threatened in communes.

Yearly deficits are totalled in April.

Dirty overcoats are drycleaned at laundries.

Labor politics is followed in Congress.

Honest citizens are married in courtrooms.

*Fancy silverware is handled in London.

WORD BOUNDARY TEST

Jim remembered that intrigues were engaging.

Sam discovered that marines are repugnant.

Hugh determined that machines are impressive.

Tests established that fatigue is depressing.

We asserted that cigars are repulsive.

Dan insisted that neglect is infernal.

Sue protested that Marie was seductive.

Jane reported that events were exciting.

John acknowledged that ballet is appealing.

Tom suspected that guitars are expensive.

*Einstein postulates that sunlight is yellow.

ACCENT PATERN TEST

Covert intentions are portrayed in cartoons.

Relaxed surroundings are preferred for croquet.

Congealed tobacco was removed from spittoons.

Repaired equipment was required for hotels.

Bizarre concoctions were prepared for dessert.

Reserve battalions were deployed for defense.

Preferred investments were obtained in Brazil.

Arcane adornments were procured for boutiques.

Verbose pronouncements were rehearsed for Truffaut.

Antique gazebos were received from Milan.

*Ancient pyramids were ransacked for plunder.

CONTROL TEST FOR TYPE IV

Roy pretended that bassoons are delightful.

She reflected that caffein is destructive.

Bill decided that alarms are important.

Mike responded that revenge was enticing.

Ted disputed that racoons are malicious.

Jill recorded that debates are essential.

Max accepted that constraints are restrictive.

Fred imagined that regimes are judicious.

James reported that reviews are instructive.

Bruce suggested that manure is offensive.

*They concluded that arrests are insulting.

SYNTAX TEST

Children memorize that strangers are evil.

Louis telegraphed that China was stormy.

Timmy verbalized that oceans are salty.

Lenin advocates that comrades be loyal.

Carter emphasized that travel is stressful.

Fairness militates that ballots be secret.

Singers vocalize that people are lonely.

Merchants calculate that profits are plenty.

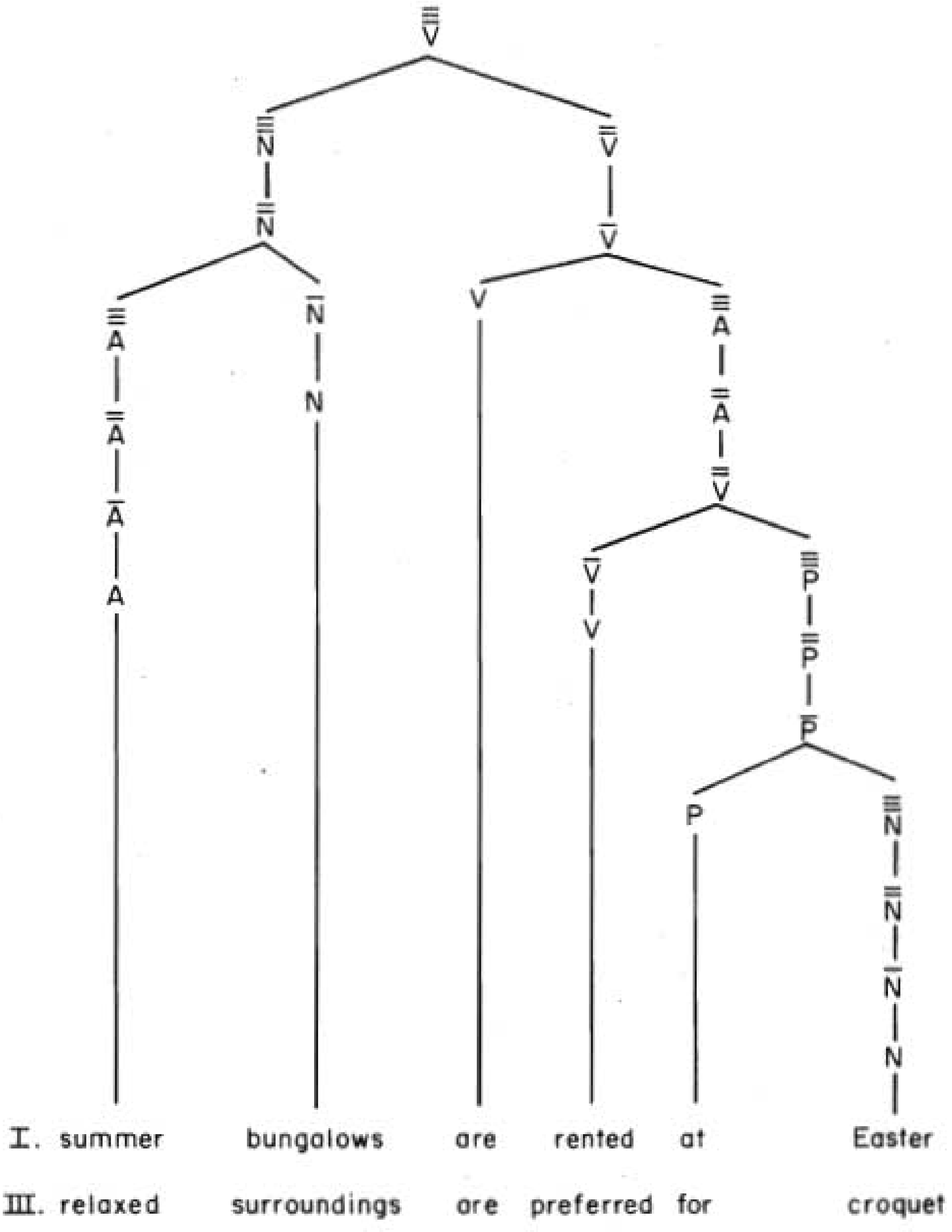
Newton theorized that motion is constant.

Diets stipulate that sugar be banished.

*Pastry bakeries are opened at midnight.

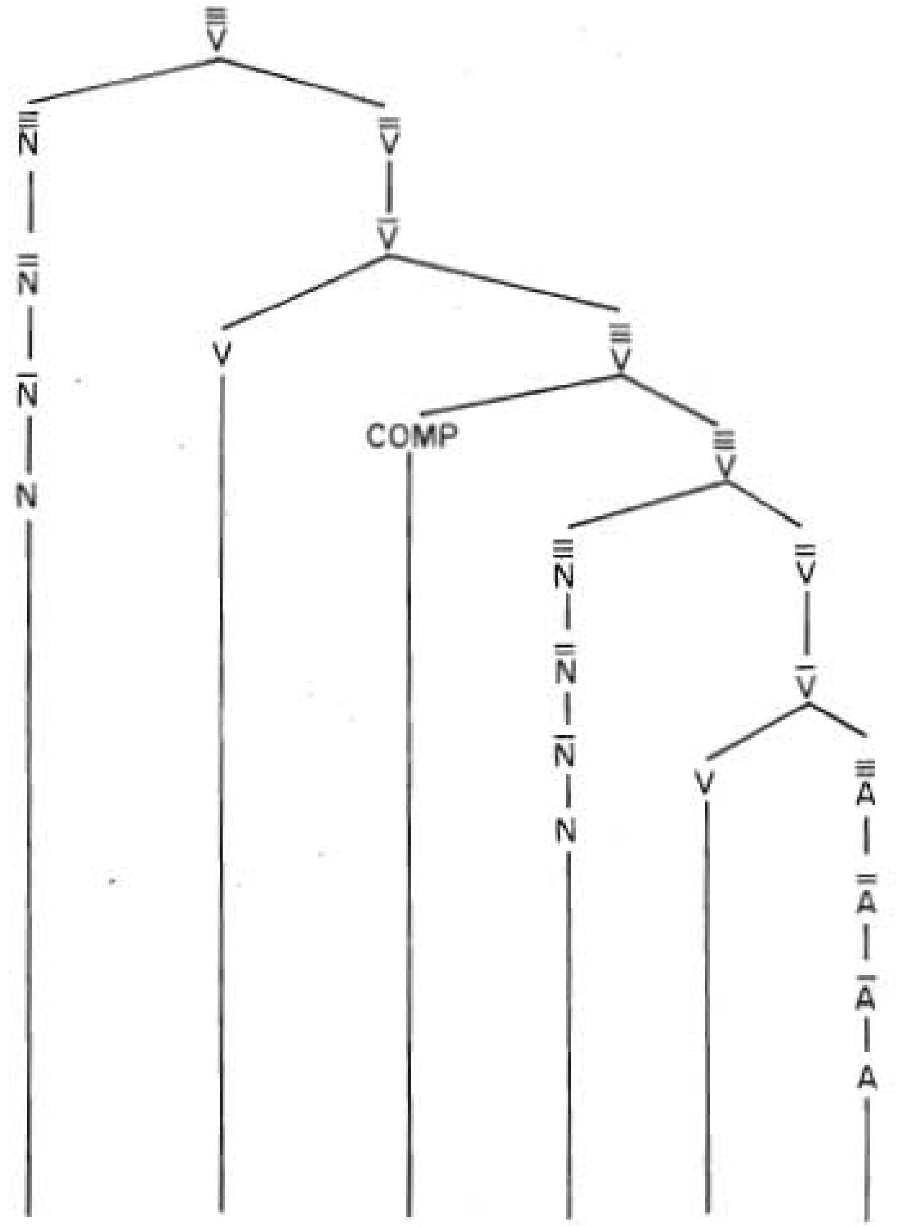
They are hāppy̆ pēoplĕ
They are prĕcīse acc̆oūnts

They are cōmbăt sōldiĕrs
They are tērrĭbl̆e to hēar



I. summer bungalows are rented at Easter

III. relaxed surroundings are are preferred for croquet



II. studies demonstrate that monkeys are clever

IV. he conceded that sardines are delicious

I.	summer	bungalows	are	rented	at	Easter
	— ◡	— ◡ ◡	◡	— ◡	◡	— ◡
III.	relaxed	surroundings	are	preferred	for	croquet
	◡ —	◡ — ◡	◡	◡ —	◡	◡ —

```
II.  studies  demonstrate  that  monkeys  are  clever
     –  ∪     –   ∪   ∪    ∪     –   ∪    ∪    –  ∪

IV.  he  conceded  that  sardines  are  delicious
     –   ∪  –  ∪   ∪     ∪   –     ∪    ∪ –  ∪
```

Some Perceptual Dependencies in Speeded Classification of Vowel Color and Pitch

Thomas D. Carrell, Linda B. Smith, and David B. Pisoni

Department of Psychology
Indiana University
Bloomington, Indiana 47405

Running head: Dependencies Between Vowel and Pitch

Abstract

Wood (1975) suggested that information specifying consonants is dependent on the earlier processing of pitch. Is vowel information also dependent on the prior processing of pitch? In contrast to the results obtained with consonants, Kuhl (1975, 1976) has reported that infants responded selectively to differences in vowel color when pitch varied but not to differences in pitch when vowel varied. Also, Miller (1978) has reported that adults show mutual, symmetric interference effects in speeded classification between vowel color and pitch. How can these results be reconciled? In this study adults' judgments of vowel color and pitch were examined in a reaction time task involving several levels of discriminability with both isolated vowels and vowels in CV syllables. The overall results were consistent with both Kuhl's and Miller's earlier data but refine our understanding of the interaction between acoustic dimensions by showing that vowel identification is indeed dependent on the processing of pitch information. Such an interaction, however, is only evident when one examines processing dependencies across a range of stimulus values for each dimension. The present findings provide new insights into the nature of processing dependencies and the methods by which such dependencies can be studied experimentally.

Some Perceptual Dependencies in Speeded Classification of Vowel Color and Pitch*

Nearly all work in speech perception is based on the assumption that several stages of processing interact to decode the acoustic waveform into a sequence of discrete units. The precise nature of these processing stages is not clear at the present time but the existence of an auditory level and a phonetic level of processing has been widely accepted in the literature on speech perception. The notion is that the auditory level operates on all acoustic information whereas the phonetic level operates only on information specific to the analysis of speech (see for example, Studdert-Kennedy, 1976).

Wood and his associates have attempted to distinguish between these two levels of processing via a speeded selective attention task (Wood, Goff, and Day, 1971; Day and Wood, 1972; Wood, 1975). In one of their experiments, subjects were presented with synthetic speech sounds that varied on auditory and phonetic dimensions and were asked to classify the sounds according to one or the other dimension (Wood, 1975). They found that subjects were able to classify syllables according to pitch unaffected by orthogonal variation in place of articulation (in this case /bae/ vs. /gae/). However, the subjects' judgements of place of articulation were slowed by orthogonal variation in pitch. This pattern of reaction time performance strongly implies an asymmetric dependency in the processing of pitch information and the processing of stop consonant information. Such a finding is consistent with the idea that the processing of pitch is independent of the processing of place of articulation whereas the processing of place of articulation depends on the possibly prior analysis of pitch.

Is pitch information, in general, independent of and prior to the analysis of the acoustic cues specifying all phonemes? There are several reasons to expect that the pattern of interaction between pitch and phoneme would be very different when the phoneme is a vowel rather than a stop consonant. The acoustic cues specifying vowels in English are very different from those specifying stop consonants. Vowels have a much slower spectral rate of change than do stop consonants. And, vowels and consonants have been shown to be processed differently in short-term memory. For example, the immediate recall of vowels show modality, recency, and suffix effects whereas consonants do not (Crowder, 1971; Crowder, 1973 a,b). In addition, it is often found that vowels are perceived continuously and that stop consonants are perceived categorically, although this outcome depends on the manner in which one tests for categorical perception (Pisoni, 1971, 1974, 1975).

A recent study by Miller (1978) addressed some questions regarding processing interactions between vowel and pitch identification more directly. In this study essentially the same experimental paradigm was used as in the earlier experiments by Wood (1975). Subjects were presented with synthetic speech sounds that varied on auditory (high versus low F0) and phonetic (/ba/ versus /bae/) dimensions. As in Woods study, subjects were required to classify these sounds according to one or the other dimension. Miller reported a mutual, symmetric interference effect. Specifically, subjects' classification time for pitch was increased by orthogonal variation of the vowel color and their classification time for syllables according to vowel color was increased by orthogonal variation of the pitch. Furthermore, the interference created by orthogonal variation of the unattended dimension was equal for both vowel and pitch identification tasks. Apparently, the processing interactions between pitch and stop consonant compared to that between pitch and vowel color are different. Precisely how they are different remains to be explored.

An experiment by Kuhl (1975) also provides information pertaining to the interaction between the perceptual dimensions of pitch and vowel color. In a sucking-habituation paradigm, Kuhl found that infants' discrimination of vowel color was unaffected by variation in pitch, but their discrimination of pitch was disrupted by variations in vowel color. This pattern suggests an asymmetric dependency between pitch and vowel color that is the exact reverse of that found by Wood between pitch and consonant. These results are quite surprising both in light of Miller's (1978) mutual symmetric results and because it seems quite unlikely that the determination of pitch is dependent on the prior or parallel processing of vowel color.

It is possible that the asymmetric interference effects reported in Kuhl's study could result from asymmetries in the discriminability of the two underlying perceptual dimensions

(Garner and Morton, 1969). For example, classification according to vowel color may not be imparied by variation in pitch if the difference between the two vowels is very large and therefore highly salient or if the difference between the two pitches is too small to produce a measurable effect since it is nondiscriminable. Because the two dimensions in Kuhl's study were not matched for discriminability, as far as we know, such a difference could account for the discrepency between Miller's and Kuhl's results. On the other hand, the difference could be due to the differences in the perception of vowels in isolation (as in Kuhl's study) versus the perception of vowels in a dynamic CV context (as in Miller's study). The two experiments reported in the present paper were undertaken to examine these explanations as well as to explore the importance of relative stimulus discriminability for perceptual interactions in speeded classification tasks.

In both experiments the perceptual interactions between pitch and vowel color were studied over a wide range of stimulus discriminabilities. This was done in order to determine how the interference effects due to processing dependencies were related to the interference effects due to differences in overall discriminability of the dimensions under study. Two experiments were carried out since there was evidence in the literature that vowels in isolation are perceived quite differently than vowels embedded in syllables. A speeded classification task was used with adult subjects as in previous studies (see Garner, 1970; Eimas et al., 1978). The stimuli consisted of isolated vowels in

Experiment 1 and vowels in a CV syllable context in Experiment 2. These experiments were designed to systematically examine the overall processing dependencies between the perception of vowel color and pitch.

Experiment 1

Method

Subjects. Seventy-eight Indiana University students were paid $3.00 per hour or were given experimental credit as part of a course requirement for their participation in this experiment. They were all native speakers of English. None of the subjects reported a previous history of speech or hearing disorder at the time of testing.

Stimuli. Five vowels at three possible pitches served as stimuli in the first experiment. Differences in vowel quality were based on a stimulus contiuum generated earlier by Pisoni (1971) that ranged perceptually from /i/ to /I/ to /ɛ/ in 13 approximately equal logarithmic steps. In this continuum /i/ is vowel number 1, /I/ is vowel number 7, and /ɛ/ is vowel number 13. Vowels between these endpoints also fall perceptually between /i/, /I/, and /ɛ/. The vowels chosen for the present experiment were stimulus numbers: 1, 3, 6, 7, and 13. Vowels 1 and 13 were used in the "Large Vowel" conditions, Vowels 1 and 7 were used in the "Intermediate Vowel" conditions, and Vowels 3 and 6 were used in the "Small Vowel" conditions. Vowel 3 is generally identified as an /i/ and vowel 6 is generally identified as an /I/ by naive subjects (Pisoni, 1971). The exact formant frequencies of these vowels are shown in Table 1.

Insert Table 1 about here

The F0 of each vowel started at one of three possible frequencies: 145 Hz, 130 Hz, or 70 Hz. In each case the F0 dropped 25 percent linearly from the onset of the vowel to the offset of the vowel. This pitch contour was implemented merely to increase the naturalness of the speech stimuli. The exact values for the F0 of each vowel are also specified in Table 1. The three pitches are numbered 1 to 3, from highest to lowest. The contrast between pitch 1 and pitch 3, a difference of 75 Hz, was used in the "Large Pitch" difference conditions and the contrast between pitch 1 and pitch 2, a difference of 15 Hz, was used in the "Small Pitch" difference conditions.

Each vowel was 150 msecs in duration with rise and fall times of 20 msecs. Stimuli were presented at approximately 80 dB SPL on TDH-39 headphones after being low-passed at 4.9 KHz. They were synthesized on a version of the Klatt software synthesizer (Kewley-Port, 1978) as implemented on a PDP 11/05 in the Speech Perception Laboratory at Indiana University in Bloomington. All stimuli were equated on the other parameters except for those specifically varied for this experiment.

Procedure. Five experimental conditions were presented to five different groups of subjects. Each group of subjects participated in only one condition. The conditions may be characterized by the degree of discriminability of the values of

Table 1

Parameter Values for Vowel Stimuli used in Experiments 1 and 2

Vowel Number	IPA Symbol	Alphanumeric Label	Pitch Contour	F1	F2	F3
1	/i/	EE1	145-116	270	2290	3010
		EE2	130-104	270	2290	3010
		EE3	70-56	270	2290	3010
3	/i✓/	EEL1	145-116	298	2226	2902
		EEL2	130-104	298	2226	2902
		EEL3	70-56	298	2226	2902
6	/I^/	IHH1	145-116	353	2103	2719
		IHH2	130-104	353	2103	2719
		IHH3	70-56	353	2103	2719
7	/I/	IH1	145-116	374	2070	2666
		IH2	130-104	374	2070	2666
		IH3	70-56	374	2070	2666
13	/ε/	EH1	145-116	530	1840	2480
		EH2	130-104	530	1840	2480
		EH3	70-56	530	1840	2480

pitch and vowel color as follows: (A) Large Vowel difference - Small Pitch difference, (B) Intermediate Vowel difference - Small Pitch difference, (C) Large Vowel difference - Large Pitch difference, (D) Small Vowel difference - Small Pitch difference and (E) Small Vowel difference - Large Pitch difference.

Subjects in each of the five conditions received three blocks of trials consisting of two control blocks and one focusing block where the target dimension was vowel color. They also received three blocks consisting of two control blocks and one focusing block where the target dimension was pitch. In the control blocks, the two stimuli differed only on the relevant dimension (Vowel or Pitch). The irrelevant dimension was fixed at one level for the Control 1 blocks and at the second level for the Control 2 blocks. In the Focusing blocks, the four stimuli were composed of all combinations of two values of vowel and two values of pitch. The resulting six test blocks which were presented to each subject are detailed in Table 2.

Insert Table 2 about here

The actual values of levels 1 and 2 depended on which of the four conditions a particular group of subjects was assigned to. For example, in the Large Vowel difference - Small Pitch difference condition, the vowels used were: EE1, EE2, EH1, and EH2. However in the Large Vowel difference - Large Pitch

Table 2

Arrangement of Stimuli for Each Block

Target Dimension	Control 1	Control 2	Focusing
			V_1P_1
Vowel	V_1P_1	V_1P_2	V_2P_1
	V_2P_1	V_2P_2	V_1P_2
			V_2P_2
			V_1P_1
Pitch	V_1P_1	V_2P_1	V_2P_1
	V_1P_2	V_2P_2	V_1P_2
			V_2P_2

difference condition, the vowels used were: EE1, EE3, EH1, and EH3 (see Table 1).

Each block consisted of 64 trials. The stimuli in each block were presented equally often in a pseudo-random order. The ordering of the blocks was counterbalenced with by means of a latin square design.

The subjects were given general instructions at the beginning of the experiment which included information that the trials would be self-paced and that the more quickly they responded, the sooner they would be finished. At the beginning of each task subjects were informed of the relevant dimension and again encouraged to respond as rapidly and as accurately as possible.

The entire session consisting of instructions and 6 blocks of 64 trials took from 40 to 50 minutes. Pilot data which was divided into quartiles indicated that practice effects leveled off after the first quartile. Therefore, the first 16 trials in each block were considered as practice trials and were not used in further analyses of the data reported here.

Results and Discussion

Mean reaction times in the five conditions are shown in Figure 1 and the specific values are given in Table 3. These means are based on correct responses only and responses over 1.75 secs were counted as errors.

Insert Table 3 and Figure 1 about here

The two control blocks for each control condition were not significantly different from each other. Thus, each pair was collapsed into one measure, both for display in Figure 1 and for subsequent statistical analysis.

A three-way factorial analysis of variance was carried out on the reaction times in each condition (subject by task by dimension). The two levels of task were control and focusing and the two levels of dimension were vowel color and pitch. In some cases the analysis showed no difference between two means. Repeated measures t-tests were then used in such cases to argue for acceptance of the "null hypothesis" that there was indeed no difference between means. The t-test was chosen simply because it is a very sensitive test for a difference between means and its repeated use would only make it more likely that a difference between means would be detected.

Percent correct scores were also calculated from the data. The overall error rate averaged 5 percent suggesting the possibility of a speed accuracy tradeoff. However, further analyses revealed that the percent correct data taken alone yielded essentially the same results as the reaction time data and therefore a speed accuracy tradeoff was not likely.

Panel A in Figure 1 illustrates the pattern of reaction times when subjects identified stimuli that varied greatly in

Table 3

Mean Reaction Times (msec.) for all Conditions in Experiment 1

Stimulus Condition	Control	Focusing
(A) Large Vowel-Small Pitch		
Pitch ID	445	595
Vowel ID	284	284
(B) Intermediate Vowel-Small Pitch		
Pitch ID	393	500
Vowel ID	320	352
(C) Large Vowel-Large Pitch		
Pitch ID	292	348
Vowel ID	292	340
(D) Small Vowel-Small Pitch		
Pitch ID	402	458
Vowel ID	391	420
(E) Small Vowel-Large Pitch		
Pitch ID	305	337
Vowel ID	354	415

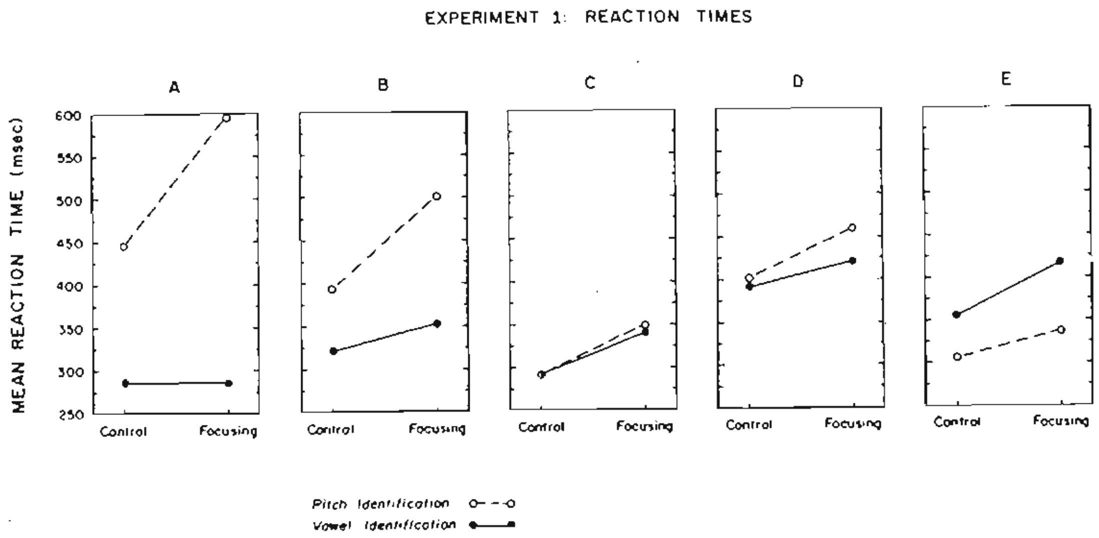


Figure 1. Mean reaction times for Large Vowel - Small Pitch, Intermediate Vowel - Small Pitch, Large Vowel - Large Pitch, Small Vowel - Small Pitch, and Small Vowel - Large Pitch conditions in Experiment 1.

vowel color but only minimally in pitch. As apparent in the figure, pitch identification was much slower than vowel identification (F (1, 11) = 29.67, $p < 0.0002$). The other main result in this condition was that irrelevant variation in the vowel impaired pitch identification, as illustrated by a longer mean reaction time in the focusing task than in the control tasks, whereas variation in pitch did not affect vowel identification. This result showed up in the analysis as a strong interaction between task and dimension (F (1, 11) = 12.95, $p < 0.004$).

It is possible that the interaction found in the first condition was due to a floor effect on vowel identification. Perhaps there was a difference between the control and focusing tasks but the subjects were already responding as rapidly as possible. In the second condition, the discriminability of the vowel was therefore reduced, in part, to examine this possibility further.

Mean reaction times for the Intermediate Vowel-Small Pitch condition are illustrated in Panel B of Figure 1. Pitch and vowel identification were closer in discriminability in this condition, as evidenced by the smaller difference in the control reaction times. However, an asymmetric pattern of interference was still obtained, as indicated by the interaction between task and dimension (F (1, 17) = 6.33, $p < 0.02$). This result indicated that the asymmetric dependency that was found in the first condition, shown in Panel A, was not entirely due to a floor effect.

The pattern of reaction times in the first two conditions of Experiment 1 have been shown to be consistent with each other. It should also be noted that these results are, at first glance, quite similar to those reported by Kuhl (1976) in her earlier investigation of infant perception.

The results of the Large Vowel - Large Pitch condition are illustrated in Panel C. In this condition a mutual symmetric pattern of interference can be observed. First, the discriminability of the pitch and vowel were matched as shown by the similarity of the vowel and pitch control reaction times ($\underline{t}$ (11) = .01, $\underline{p}$ < 0.9). Second, consistent with a hypothesis of mutual interference, the variation of the irrelevant dimension interferred with the identification of the relevant dimension ($\underline{F}$ (1, 11) = 13.37, $\underline{p}$ < 0.004). And third, consistent with the hypothesis of symmetric interference, there was no interaction between task and dimension ($\underline{F}$ (1, 11) = .06, $\underline{p}$ < 0.81).

It is conceivable that the pattern of results shown in Panel C of Figure 1 could have been influenced by a floor effect in the control tasks for both vowel and pitch. There may have been a discriminability difference not reflected in the reaction time scores. To test this possibility, the discriminabilities of both vowel and pitch were reduced in the Small Vowel-Small Pitch condition. The vowel and pitch discriminabilities were again equated in the control conditions ($\underline{t}$ (17) = .44, $\underline{p}$ < 0.7) at this lower discriminability level and the results are summarized in Panel D. The analysis of variance showed the same pattern of results in the Small Vowel-Small Pitch condition as it did in the

Large Vowel-Large Pitch condition. The main effect of task was significant ($\underline{F}$ (1, 17) = 9.07, $\underline{p}$ < 0.008). The task by dimension interaction was not significant ($\underline{F}$ (1, 17) = .77, $\underline{p}$ < 0.39). The pattern of both results indicates a symmetric pattern of interference over the two dimensions.

The pattern of reaction times illustrated in Panels C and D, in which the discriminabilities of the stimuli were equated, indicated a mutual symmetric interference between vowel and pitch. These results are therefore consistent with Miller's (1978) earlier findings with adults. However, she used CV stimuli rather than isolated vowel stimuli.

The final condition in Experiment 1 was the Small Vowel-Large Pitch condition. The results are illustrated in Panel E. The discriminability of the pitch stimuli was greater than the discriminability of the vowel stimuli in this condition as shown by the differences in the control task reaction times. A strong main effect of dimension supported this difference ($\underline{F}$ (1, 17) = 21.02, $\underline{p}$ < 0.0003). Since there was a main effect of task ($\underline{F}$ (1, 17) = 12.04, $\underline{p}$ < 0.003) coupled with no interaction between dimension and task ($\underline{F}$ (1, 17) = 3.13, $\underline{p}$ < 0.1), a mutual symmetrical interference pattern between the dimensions can be inferred. However, these results indicated that interference was <u>not</u> completely determined by overall discriminability. Specifically, when vowel was less discriminable than pitch an asymmetric interference pattern was not obtained. Recall from the first two conditions that when vowel was more discriminable than pitch an asymmetric interference pattern was obtained.

Based on these analyses, the results for Experiment 1 indicate that when vowel was relatively more discriminable than pitch, an asymmetric interference pattern was found. On the other hand, when vowel and pitch discriminabilities were equated or when pitch was relatively more discriminable than vowel a symmetric mutual interference pattern was found.

To show the relationship between discriminability and interference relation more clearly we have replotted these results in Figure 2. On the ordinate we have displayed a measure of interference in milliseconds. This measure is the difference between the focusing and control reaction times for each of the five experimental conditions. The abscissa represents a measure of the discriminability of the two dimensions. It is the difference between the pitch and vowel controls for each of the five conditions. At the 0 point on the X axis the reaction time was equal for both pitch and vowel control groups. As the values increase to the right of 0, the relevant stimuli (a solid line for the vowel relevant conditions and a dashed line for the pitch relevant conditions) are more discriminable and as the numbers decrease to the left of 0, the relevant stimuli are less discriminable.

Insert Figure 2 about here

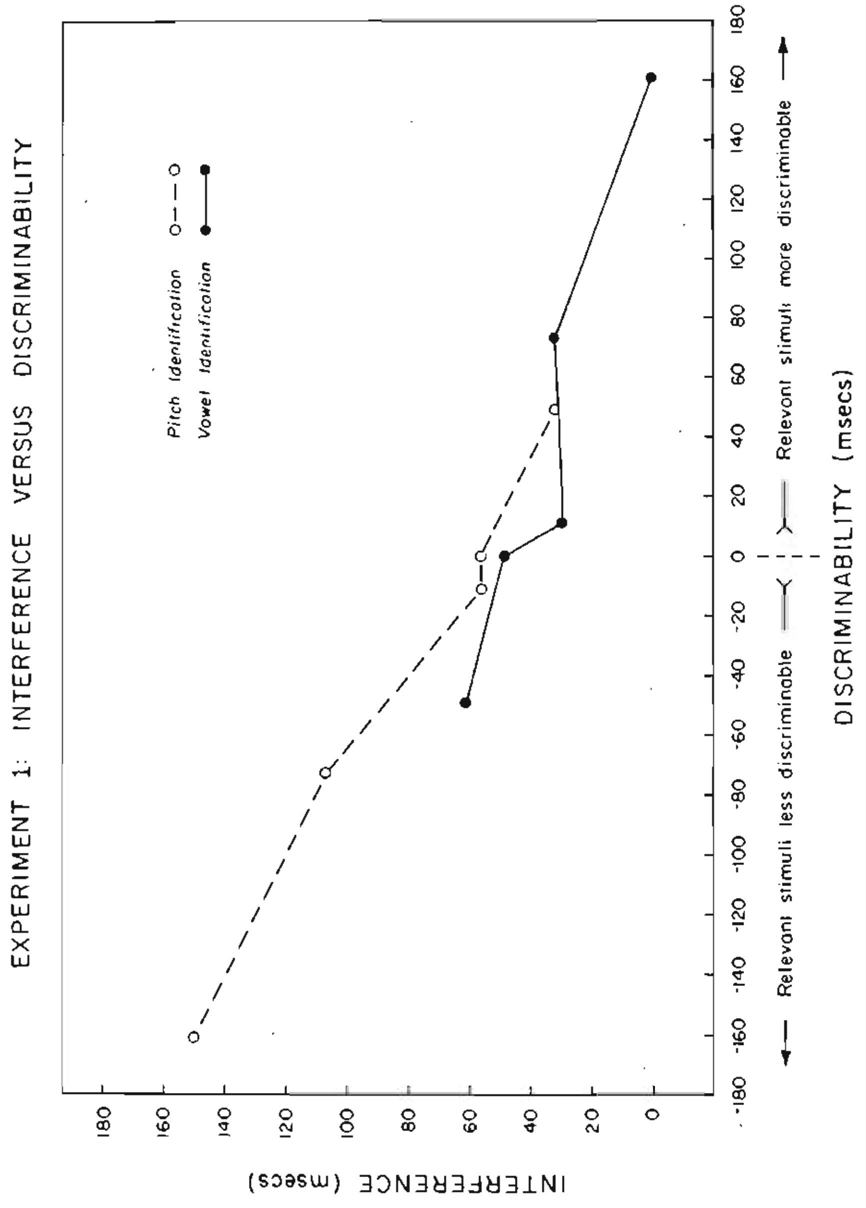


Figure 2. Interference versus discriminability for isolated vowels.

Even though the abscissa is labeled in msec the difference between the two control conditions is really a function of the underlying discriminability, and not a strict metric of discriminability. Nevertheless, if we assume that the X axis is at least monotonically increasing with discriminability, it is possible to examine the degree of observed interference as a function of discriminability. For example, it is apparent that the vowel dimension was relatively more discriminable than the pitch dimension. This is indicated by the vowel function's general right shift compared to the pitch function. Further support for this finding was was obtained from pilot data which indicated that even with a much greater pitch difference (i.e. 135 Hz in the Large Pitch difference condition rather than 75 Hz as used in the present experiments) pitch discriminability was not increased compared to vowel discriminability.

Figure 2 also suggests that, across several levels of discriminability, an asymmetric interference pattern exists between vowel color and pitch. Specifically, irrelevant vowel variation appears to interfere with pitch identification slightly more than irrelevant pitch variation interferes with vowel identification. This relation is indicated in Figure 2 across the range of discriminabilities which overlap for vowel and pitch identification; that is from about -40 msecs to +50 msecs on the abscissa. Asymmetric interference was not found, however, in the analyses of variance for the equal discriminability conditions. In spite of the findings for the equal discriminability conditions, the overall functions showing the relation between discriminability and interference in Experiment 1 displayed a consistant although small asymmetry across conditions suggesting that the asymmetric interference pattern should not be discounted as a spurious finding. That is,irrelevant vowel variation does appear to interfere with pitch identification slightly more than irrelevant pitch variation interferes with vowel identification.

Experiment 2

In Experiment 1 the stimuli were presented as isolated vowels. It has often been suggested that isolated vowels are "unnatural" and therefore not perceived as speech. A redundant consonant (/b/) was appended to the beginning of each stimuli in Experiment 2 in order to examine the interference between vowel color and pitch when the stimuli were more "speech-like."

Method

Subjects. The subjects in this experiment were 102 Indiana University students. They were paid $3.00 per hour or were given experimental credit for their participation. They were all native speakers of English. None of the subjects reported a previous history of speech or hearing disorder at the time of testing.

Stimuli. The only difference between the stimuli in Experiment 1 and Experiment 2 was that the vowels were embedded in a dynamic CV syllable context rather than being placed in isolation. A /b/ was appended to the beginning of each stimulus so that they ranged perceptually from a /bi/ to a /bI/ to a /bɛ/. There were no other differences between the stimuli in the two experiments.

Procedure. Again, five experimental conditions were presented to five different groups of subjects. Each group of subjects participated in only one condition. The conditions may be characterized as follows: (A) Large Vowel difference-Small Pitch difference, (B) Intermediate Vowel difference-Small Pitch difference, (C) Large Vowel difference-Large Pitch difference, (D) Small Vowel difference-Small Pitch difference, and (E) Small Vowel difference-Large Pitch difference. The subjects were run in the same manner and with the same apparatus as described earlier for Experiment 1.

Results and Discussion

Mean reaction times in the five conditions are shown in Figure 3 and Table 4. The data were analyzed via an analysis of variance in the same manner as described for Experiment 1.

Insert Table 4 and Figure 3 about here

The Large Vowel-Small Pitch condition and the Intermediate Vowel-Small Pitch condition (illustrated in Panels A and B respectively) reveal an asymmetric interference pattern in which irrelevant vowel variation interfered with pitch identification. This relation was supported by the strong task by dimension interactions in Panel A ($\underline{F}$ (1, 17) = 10.85, $\underline{p}$ < 0.004) and in Panel B ($\underline{F}$ (1, 17) = 25.99, $\underline{p}$ < 0.00005). In both panels, vowel was much more discriminable than pitch as shown by the highly significant main effects of dimension in Panel A

Table 4

Mean Reaction Times (msec.) for all Conditions in Experiment 2

Stimulus Condition	Control	Focusing
(A) Large Vowel-Small Pitch		
Pitch ID	407	513
Vowel ID	290	278
(B) Intermediate Vowel-Small Pitch		
Pitch ID	356	443
Vowel ID	299	314
(C) Large Vowel-Large Pitch		
Pitch ID	299	376
Vowel ID	293	368
(D) Small Vowel-Small Pitch		
Pitch ID	404	483
Vowel ID	411	502
(E) Small Vowel-Large Pitch		
Pitch ID	298	335
Vowel ID	393	571

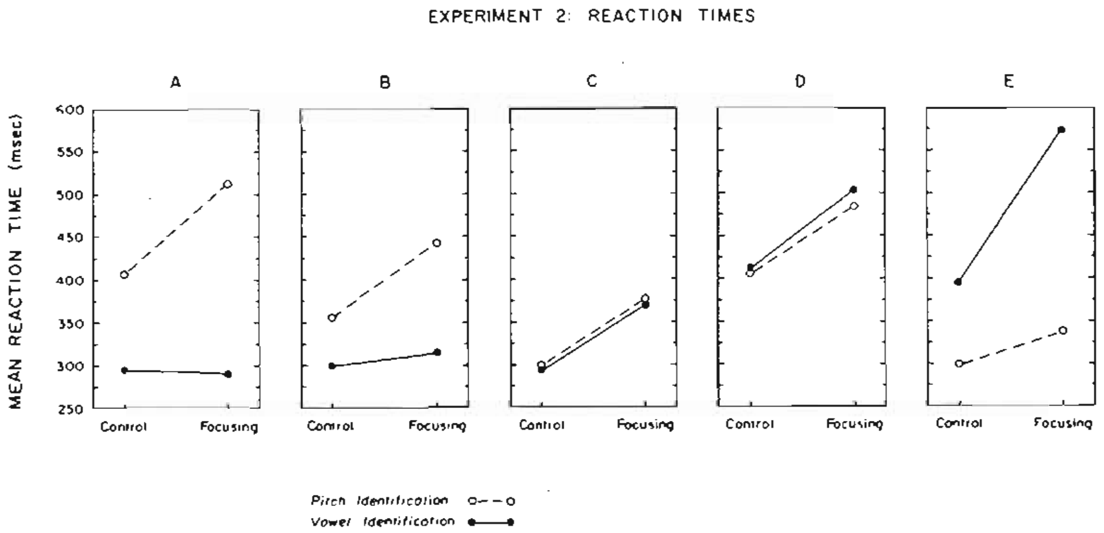


Figure 3. Mean reaction times for Large Vowel - Small Pitch, Intermediate Vowel - Small Pitch, Large Vowel - Large Pitch, Small Vowel - Small Pitch, and Small Vowel - Large Pitch conditions in Experiment 2.

(F (1, 17) = 54.01, $p < 0.00001$) and in Panel B (F (1, 17) = 35.53, $p < 0.00001$).

In Panels C and D, representing the Large Vowel-Large Pitch and the Small Vowel-Small Pitch conditions, relative discriminabilities between vowel and pitch were equal, (t (17) = .31, $p < 0.5$ and t (23) = .22, $p < 0.5$ respectively) although the absolute level of discriminability was lower in Panel D than in Panel C. A mutual and symmetric pattern of interference was found in both conditions as indicated by significant main effects of task as shown in Panels C and D (F (1, 17) = 14.58, $p < 0.002$ and F (1, 23) = 11.14, $p < 0.003$ respectively). Further support for this interpretation is provided by the absence of an interaction in both conditions (F (1, 17) = .004, $p < 0.95$ and F (1, 23) = .13, $p < 0.73$). Thus, in these two conditions pitch identification and vowel identification showed symmetric interference effects due to irrelevant variation in the unattended dimension.

Panel E illustrates the results of the Small Vowel-Large Pitch condition, a condition in which the pitch was much more discriminable than the vowel. In this condition an asymmetric interference pattern opposite to that shown in Panels A and B was found, as indicated by the task by dimension interaction (F (1, 17) = 7.16, $p < 0.016$).

The results across all five conditions in Experiment 2 displayed a consistent pattern. When the vowel dimension was more discriminable than the pitch dimension, an asymmetric interference effect was evident. Vowel identfication was unaffected by irrelevant changes in pitch whereas pitch identification performance was much reduced by irelevant changes in vowel. Conversely, when pitch was more discriminable than vowel, an asymmetric interference effect was obtained in the opposite direction. Finally, the conditions in which the relative discriminability of vowel and pitch were matched showed a mutual symmetric interference pattern.

The relationship between discriminability and interference suggested earlier in Figure 3 has been replotted in Figure 4 so it may be seen more clearly. In contrast to Experiment 1, the ranges of relative discriminability levels across vowel and pitch were quite close in Experiment 2. Recall that in Experiment 1 the range of discriminabilities was higher (i.e. more discriminable) for vowel identification than for pitch identification.

Insert Figure 4 about here

There are two other effects shown in Figure 4 which contrast with those obtained in Experiment 1. First, changes in discriminability affected vowel identification more strongly than pitch identification. This is reflected by the steeper slope of the vowel function when compared to the pitch function shown in Figure 4. Second, pitch and vowel identification were equally affected by discriminability when they were both equally discriminable. This was clearly demonstrated by the fact that the

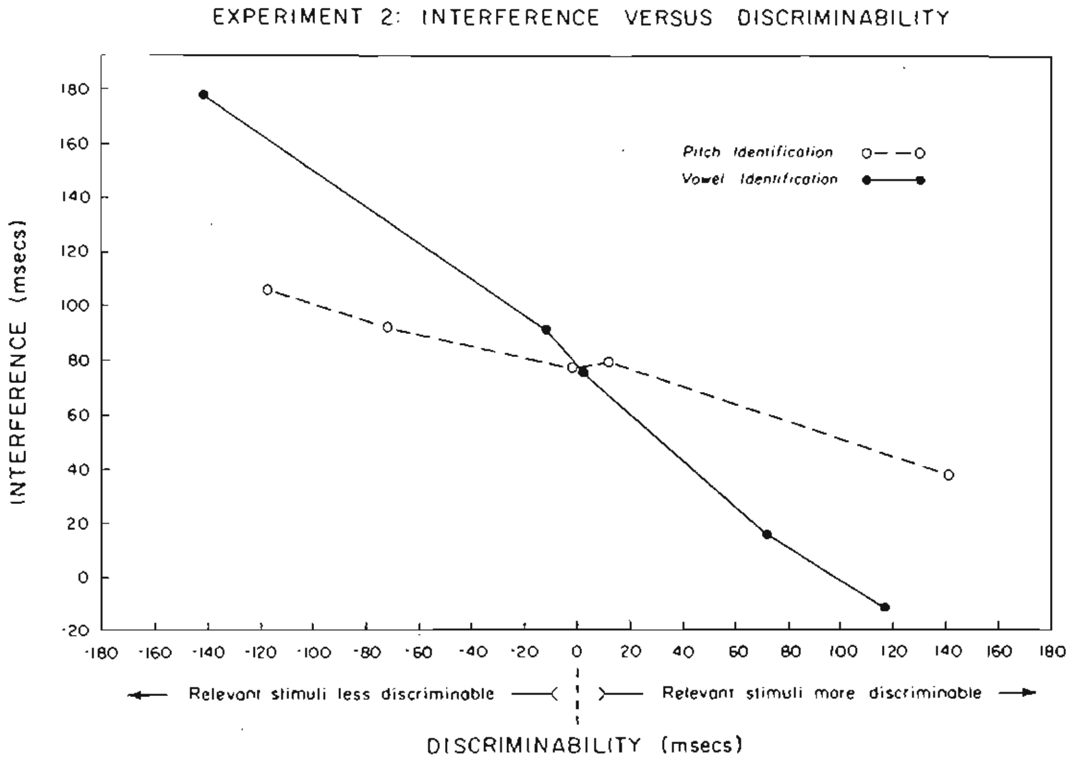


Figure 4. Interference versus discriminability for vowels in a CV context.

vowel and pitch functions intersect at a discriminability value of exactly 0 msecs in Figure 4.

To summarize, the results of Experiment 2 demonstrated that in a CV syllable context a mutual symmetric interference between vowel identification and pitch identification can be observed when relative discriminabilities are matched. However, this interference relationship could be systematically altered in either direction by changes in the discriminability of the component dimensions.

General Discussion

The overall pattern of results obtained in the present experiments supports the argument that interference effects found in speeded selective attention tasks are systematically related to several factors including relative discriminability, processing dependencies, and possibly even salience differences between the dimensions. More importantly, however, these experiments have also shown that when the dimensions are vowel color and pitch, one interference pattern was found if the stimuli are presented in isolation and another if the stimuli are embedded in a dynamic CV syllable context.

In a speeded classification task interference effects between orthogonal dimensions have been proposed to be related to relative discriminability, processing dependencies, and salience differences for the following very general reasons. First, interference may be caused because the irrelevant dimension may be so highly discriminable that it would be difficult for a subject to ignore selectively. In the reverse situation, the target dimension may have such low discriminibility that it would be difficult to detect. Second, processing dependencies may create interference because prior or parallel information from the irrelevent dimension would be necessary for recognition or identification of the target dimension. And finally, salience differences between the dimensions may cause interference because one dimension may draw the subjects attention more than the other. All of these factors apparently played a role in both of the present experiments although the clearest relationship was between relative discriminability and interference.

In Experiment 1, as the level of relative discriminability of the relevant or focused dimension was increased, the interference from the orthogonally varying dimension decreased. This discriminability versus interference tradeoff was slightly stronger for pitch identification than for vowel identification. That is, pitch was affected more by changes in discriminability than vowel was. The difference was not substantial, however, and was probably due to the ranges of discriminability covered by pitch and by vowel. The pitch function covered values of relative discriminability that were somewhat lower, in general, than those covered by the vowel function. Nevertheless, it was generally true that as relative discriminability increased the interference effects were weakened. Furthermore, when interference and discriminability were defined in terms of reaction time displayed in Figure 2, the relationship between discriminability and interference appeared to be quite linear. Since the slopes of the interference by discriminability functions were so similar for both vowel and pitch, it is reasonable to assume that vowel quality was being processed in the same manner as pitch, at least in this experimental context.

The fact that pitch identification tasks covered a range of discriminabilities which was lower than the range covered by the vowel identification tasks was quite striking. However, as reported earlier in the present study, pilot data indicated that this difference was not due to the particular stimulus values chosen for presentation on each dimension. We were not able to increase the discriminability of the pitch dimension compared to the vowel dimension either by increasing the difference in fundamental frequency or by decreasing the difference in vowel color. These findings indicated that vowel color was generally more discriminable than pitch when the stimuli were presented as isolated sustained vowels.

Experiment 1 also showed that when the discriminabilities of the two dimensions were closely matched, a mutual interference pattern was observed. This pattern of interference can be explained in terms of processing dependencies. That is, the perception of one dimension would be at least partially dependent on the processing of attributes of the other dimension. Recall, however, that there was a slight indication that the mutual interference pattern was not completely symmetric. Irrelevant variation of vowel color interfered with pitch identification slightly more than irrelevant variation of pitch interfered with vowel identification. Since this small effect was found at all levels of discriminability that vowel color and pitch had in common, it is probably safe to assume that the asymmetry was not due to discriminability effects per se and was more likely due to a processing dependency or to a salience difference between the two dimensions under study.

The present results fit nicely with the earlier data of Kuhl (1975) as far as they went. In our isolated vowel conditions in which the relative discriminabilities were matched, we did find a slight, but nonsignificant, asymmetric interference pattern in the same direction as Kuhl found in her earlier infant experiment using quite a different methodology. However, when the vowel discriminability was increased in comparison to the pitch discriminability, we obtained an interference pattern which exactly duplicated the pattern she had found. Specifically, orthogonal variation in vowel color interferred with the detection of pitch but orthogonal variation in pitch had no effect on the detection of vowel color. Kuhl explained her results in terms of salience, arguing that vowel quality is somehow intrinsically more salient to infants than pitch because vowel information is phonetic whereas pitch information is not. Based on our present findings and the existence of a large number of tone languages in which pitch is phonemic, this account must clearly be incorrect. The present analysis indicates that in Kuhl's experiment, as in two of the conditions in our first experiment, a high level of vowel discriminability could easily account for the observed asymmetric interference patterns that we obtained. It would seem reasonable, based on these findings, to account for Kuhl's results in a similar manner.

In Experiment 2 a discriminability versus interference tradeoff was also observed which was somewhat similar to the tradeoff obtained in Experiment 1. As the level of relative discriminability of the relevant dimension was increased, the interference from the orthogonally varying dimension decreased. In contrast to Experiment 1, however, there were two distinct patterns of interference, one for vowel identification and another for pitch identification. The relation between discriminability and interference appeared to be linear for both dimensions although the slope was much steeper for vowel identification. These reasults indicated that vowel color was much more affected by changes in relative discriminability than pitch. The results also showed that when vowels are embedded in a CV syllable, processing of vowel color appears to be quite different than processing of pitch.

In previous investigations of auditory processing dependencies that used speeded selective attention tasks, it has been assumed that if the discriminabilities of the component dimensions were matched, an asymmetric interference pattern indicated an asymmetric processing dependency and a mutual interference pattern indicated a mutual processing dependency. One exception to this trend was Wood (1975) who attempted to correlate interference patterns with cortical evoked potential data in order to make the leap from interference to processing dependencies. Unfortunately, evoked potential data of this kind is very hard to interpret and can only be used as weak evidence for the specific processing dependencies Wood was studying. The assumption just described which simply ties interference patterns to processing dependencies is specifically not made in the present analysis. Processing dependencies are inferred by a comprehensive exploration of interference patterns across many levels of discriminability. In fact, our comprehensive analysis of Experiment 2 indicates different processing dependencies than would be predicted by traditional analyses of speded classification data.

In spite of the mutual symmetric interference that was found between vowel color and pitch in Experiment 2, the functions plotted across conditions showed an overall asymmetric dependency. Vowel identification was quite dependent on pitch information because small changes in the relative discriminability between vowel and pitch produced relatively large changes vowel identification latencies. Pitch identification, on the other hand, was not nearly as dependent on vowel information because large changes in the relative discriminability between vowel and pitch made only relatively small changes in pitch identification latencies. This interpretation of the data shows that vowel identification is much more dependent on pitch information than pitch identification is dependent on vowel information.

The slope of the interference versus discriminability function in the pitch condition of Experiment 2 was similar to the slopes found for the pitch and vowel identification in Experiment 1. This similarity shows that, on some level, pitch in a CV syllable was processed similarly to both dimensions in an isolated vowel. Conversely, identification of vowel color in a CV syllable was processed quite differently than either dimension in an isolated context.

Miller's (1978) results also fit nicely with the present analysis although her interpretation assumed that patterns of interference were due exclusively to processing dependencies between the dimensions. Recall that in her experiment, CV syllables varying in vowel color (/a/ and /ae/) and in pitch were presented for speeded classification. The dimensions were equated for relative discriminability in the usual manner and a symmetric mutual dependency was obtained. This finding corresponds precisely to the results of the conditions examined in Experiment 2 in which the dimensions were matched for relative discriminability. Miller stated that the mutual interference was caused because vowel quality was being processed at the same level as pitch. Alternatively, we would propose that there was a processing dependency in her task but it did not show up because interference was only measured at one level of relative discriminability.

Our results in the equal discriminability conditions of Experiment 2 do contrast with the interference effects observed in identification of stop consonants and pitch found by Wood (1975). He found an asymmetric pattern of interference between stop consonant and pitch, whereas we found a mutual symmetric interference pattern. The interpretation of our results did not, however, conflict substantially with Wood's interpretation. He argued that pitch was being processed at an acoustic level while stop consonants were being processed at a phonemic level. We make essentially the same argument; vowel color is processed at a level which requires the prior or parallel analysis of pitch. Furthermore, we would predict that if the processing dependency between consonant and pitch which he proposed does indeed exist, an analysis across values of relative discriminability would show a steeper consonant identification function than pitch identification function. The phonemic level, according to Wood, relies on prior acoustic analysis of the signal whereas the acoustic level has no similar reliance on the phonemic level. Although we do not use his terminology, it is clear from our results in Experiment 2 that the identification of vowel is, in fact, dependent on prior pitch information when the vowel is embedded in a CV syllable context.

Additional experiments similar to those reported here will be necessary for a better understanding of processing dependencies between component dimensions. It will be critically important to examine these relations across a wide range of discriminabilities since simply matching discriminabilities may not reveal the whole story. Experiments using other consonants, liquids, fricatives, and less commonly analyzed speech sounds will also be necessary in order to explain the complex processing dependencies that exist for speech sounds.

In the final analysis, however, the explanation for the different processing dependencies in the perception of pitch, consonant, and vowel color in speeded classification tasks may be traceable to the acoustic differences between these dimensions and their overall similarity. Given the large acoustical differences in the cues for stop consonants and vowels, different processing dependencies would seem necessary for their early perceptual analysis and recognition. However, will the information gained from a purely acoustic analysis simply generate a trivial list of attributes in need of a higher level of organization, such as phonetic or feature based, or will it be tied systematically to some physical aspect of the stimuli? Further research with stimuli varying on different dimensions and tested at different levels of discriminability will be necessary to provide a principled answer to this question and to further illuminate the complex relationships observed in processing multidimensional signals such as speech.

References

Crowder, R. G. The sound of vowels and consonants in immediate memory. Journal of Verbal Lerning and Verbal Behavior, 1971, 10, 587-596.

Crowder, R. G. Representation of speech sounds in precategorical acoustic storage. Journal of Experimental Psychology, 1973a, 98, 14-24.

Crowder, R. G. Precategorical acoustic storage for vowels of short and long duration. Perception & Psychophysics, 1973b, 13, 502-506.

Day, R. S., & Wood, C. C. Mutual interference between two linguistic dimensions of the same stimuli. Journal of the Acoustical Society of America, 1972, 52,175. (Abstract)

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. Asymmetric dependencies in processing phonetic features. Perception & Psychophysics, 1978, 23, 12-20.

Garner, W. R. The stimulus in information processing. American Psychologist, 1970, 25, 350-351.

Garner, W. R., & Morton, J. Perceptual independence: Definitions, models, and experimental paradigms. Psychological Bulletin, 1969, 72, 233-259.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. *Research on Speech Perception Progress Report No. 4*, Indiana University, Bloomington Indiana, 1978.

Kuhl, P. K. Speech Perception in early infancy: The aquisition of speech-sound categories. In S. K. Hirsh, D. H. Eldridge, I. J. Hirsh, & S. R. Silverman (Eds.), *Hearing and Davis: Essays honering Hallowell Davis*. Saint Louis: Washington University Press, 1976.

Kuhl, P. K., & Miller, J. D. Speech perception in early infancy: Discrimination of speech-sound categories. *Journal of the Acoustical Society of America*, 1975, *58*, S56.

Miller, J. L. Interactions in processing segmental and suprasegmental features of speech. *Perception and Psychophysics*, 1978, *24*, 175-180.

Ladefoged, P. *A Course in Phonetics*. New York: Harcourt Brace Jovanovich, Inc., 1975.

Nickerson, R. S. Binary-classification reaction time: A review of some studies of human information-processing capabilities. *Psychonomic Monograph Supplements*, 1972, *4*(17, Whole No. 65).

Pisoni, D. B. Information processing and speech perception. In G. Fant (Ed.), *Speech communications* (Vol. 3). New York: Wiley, 1975.

Pisoni, D. B. *On the nature of categorical perception of speech sounds*. Unpublished doctoral dissertation, University of Michigan, 1971.

Pisoni, D. B., & Lazarus, J. H. Categorical and noncategorial modes of speech perception along the voicing continuum. Journal of the Accoustical Society of America, 1974, 55, 328-333.

Singh, S., & Woods, D. R. Perceptual structure of 12 American English vowels. Journal of the Acoustical Society of America, 1970, 49, 1861-1866.

Studdert-Kennedy, M. Speech Perception. In Lass, N. J. (Ed.), Contemporary Issues in Experimental Phonetics, Springfield, Illinois: C. C. Thomas, 1974.

Wood, C. C. Auditory and phonetic levels of processing in speech perception: Neurophysiological and information processing analysis, J. Exp. Psychology: Human Percep. and Performance. 1975, 104, 1-33.

Wood, C. C., Goff, W. R., & Day, R. S. Auditory evoked potentials during speech perception. Science, 1971, 173, 1248-1251.

Footnote

*This research was supported in part by NIMH Research Grant MH-24027-05 and in part by NIH grant NS-12179-04 to Indiana University in Bloomington.

Adaptation and Contrast in the Perception of Voicing

James R. Sawusch

State University of New York at Buffalo

Buffalo, New York 14226

and

Peter Jusczyk

Dalhousie University

Halifax, Nova Scotia B3H 4J1 Canada

Short title: Adaptation and Contrast

Abstract

The results of experiments using selective adaptation with stop consonants have been interpreted in terms of auditory feature detector fatigue, phonetic feature detector fatigue and response contrast. In the present studies, both a selective adaptation procedure and a procedure involving paired comparisons between successively presented stimuli were used to sort out these explanations. A fricative-stop-vowel syllable ([sba]) was constructed using an [s], followed by 75 msec of silence, followed by a 0 msec VOT [ba]. The perceived phonetic identity of this syllable was [p] even though the spectral structure of the stop-vowel within this syllable was identical to the [ba] end of a [ba] - [p^ha] test series. As adaptors, the [sba] and [ba] endpoint syllables had identical effects. In the paired comparison procedure, the [sba] had no effect on an ambiguous test item while the [ba] caused the test item to be labelled as [p]. Results of these experiments indicate that neither response contrast nor phonetic feature detection are involved in selective adaptation effects found for a voicing stop-consonant series. Results were interpreted as supporting the position that selective adaptation effects arise at an early, auditory level(s) of processing that is responsive to the spectral overlap between adaptor and test items.

Over the last few years, the selective adaptation paradigm has been employed extensively in an effort to understand the basic processes underlying speech perception. The original impetus for applying this paradigm to the study of speech perception was the belief that there might be detector mechanisms in the brain which are sensitive to certain linguistic features. It was assumed that one way of demonstrating the existence of such detector mechanisms would be to show that repeated presentation of a feature to which a detector is sensitive would fatigue the detector and thus reduce its sensitivity to sounds containing that feature (Eimas & Corbit, 1973).

Early studies uncovered evidence for selective adaptation and attributed such effects to the existence of detectors specialized for processing speech (Eimas, Cooper & Corbit, 1973; Eimas & Corbit, 1973). However, subsequent research has raised the possibility that the existence of a level of phonetic feature detectors in neither sufficient nor necessary to account for the effects of selective adaptation to speech (Ades, 1976; Bailey, 1975; Diehl, Elman & McCusker, 1978; Sawusch, 1977a). In fact, some investigators have questioned whether selective adaptation studies provide evidence for the existence of any kind of feature detector, be they phonetic or auditory (Diehl et al., 1978; Remez, 1979; Remez, Cutting & Studdert-Kennedy, in press; Simon & Studdert-Kennedy, 1978).

Thus, at the present time, there is considerable disagreement as to how the data from selective adaptation studies are to be interpreted. At the risk of over-simplification, there appear to be three basic positions with respect to this issue. The first position treats adaptation effects as the result of feature detector fatigue at an auditory level of processing. This fatigue is governed by the spectral overlap between tha adapting stimulus and the members of the test series (Ades, 1976; Bailey, 1975; Eimas & Miller, 1978; Pisoni & Tash, 1975). According to this position, repeated presentations of the adaptor should lead to desensitization of the detector which responds to it. Consequently, when some members of the test series overlap spectrally with the adaptor, there should be an indication that after prolonged exposure to the adaptor, subjects are less sensitive to those members of the test series. On the other hand, such sensitivity changes would not be expected to occur when the adaptor does not overlap spectrally with members of the test series, even when some members of the series share the same "phonetic" label as the adaptor. Precisely this result has been reported by Sawusch (1977a) who showed that, after adaptation, significant drops occurred in the rated quality of stimuli within a phonetic category when they overlapped spectrally with the adaptor. Similar findings have also been reported by Miller, Eimas & Root (1977).

The second line of argument concerning adaptation-induced changes attributes these effects to two distinct levels of processing in speech perception (Cooper, 1975; Sawusch, 1977a; Simon & Studdert-Kennedy, 1978; Tartter & Eimas, 1975). One of these levels consists of a set of auditory feature detectors, as outlined above. The second level has been described two ways. In early work, Cooper (1975) and Tartter & Eimas (1975) described this second level as consisting of phonetic feature detectors. More recently, Sawusch (1977a) has characterized this second level of processing as an abstract, auditory one which responds to patterns or configurations of cues in the speech signal. Furthermore, he suggested that the effect of adaptation at the second level can be described as a retuning of the pattern processing mechanism rather than detector fatigue. Simon &

Studdert-Kennedy (1978) have offered a similar analysis of the selective adaptation results and described the second level as one that operates by "auditory contrast". Thus, the primary distinctions between the proposals offered by Sawusch (1977a; also Simon & Studdert-Kennedy, 1978) and by Cooper (1975; also Tartter & Eimas, 1975) are: (1) the degree of abstractness of the second level of processing (i.e., abstract auditory versus phonetic); and (2) the type of effect adaptation has at this second level (retuning versus fatigue). The evidence favoring two-level models of adaptation comes from experiments where adaptation effects have been found even in the absence of spectral overlap between adaptor and test series (Sawusch, 1977a; Ganong, Note 1).

The third position with respect to adaptation effects rejects the notion of any sort of feature-detector fatigue and instead attributes such effects solely to response contrast (Diehl et al., 1978; Diehl, Lang & Parker, in press). Diehl and his colleagues have suggested that since adaptors are typically good exemplars of a phonetic category, they may serve as a reference against which subjects judge members of the test series. Thus, rather than attributing phonetic boundary shifts to sensory fatigue, Diehl at al. claim that such shifts are the result of modifications of subjects' phonetic decision criteria (see also Elman, 1979). In support of their contention, these investigators present data from three experiments showing that boundary shifts can be produced for test items near a phonetic category boundary by merely pairing such items with good exemplars of categories involving the same phonetic dimension. Moreover, Diehl et al. demonstrated that such response contrast effects could be obtained even in cross-series tests where the exemplars differed spectrally from the test items (e.g. a [ga] exemplar can affect the labeling of test items from a [ba] - [p^{h}a] series).

At first glance, Diehl et al.'s data would seem to provide insurmountable problems for accounts based upon the notion of feature-detector fatigue, for it is difficult to see how a single presentation of a good exemplar could lead to adaptation. However, as Diehl et al. have indicated, the size of the boundary shifts which they have obtained is smaller than those usually observed in selective adaptation studies. This latter fact raises the possibility that the response contrast explanation may not be sufficient to account for the full extent of adaptation effects previously reported in the literature.

Evidence pertinent to distinguishing among these three positions has been difficult to muster since most previous research has been conducted using adaptors which did not dissociate the spectral characteristics of the stimulus from its phonetic identity. Consequently, the various accounts of adaptation typically predict the same set of results. Moreover, even those instances in which investigators have tried to dissociate the spectral and phonetic cues have proven somewhat equivocal since in some cases adaptation was obtained (Cooper, 1974; Cooper & Blumstein, 1974; Ganong, 1978; Sawusch, 1977a; Ganong, Note 1) but in other cases it was not (Ades, 1974; Bailey, 1975; Sawusch, 1977b).

The chief difficulty with many of the previous efforts to dissociate spectral and phonetic information during adaptation has been that while the adaptor shared some phonetic dimension with one end of the test continuum, its spectral characteristics were often quite discrepant from members at both ends of the continuum. For example, this is the case when one uses [sae] as an adaptor and tests for its effect on the identification of a

[bae] - [dae] series. Still, there have been some attempts to pit the spectral characterisitcs of the adaptor directly against its phonetic identity (Ades, 1974; Sawusch, 1977b; Wolf, 1978; Pisoni, Note 2). The results of these studies have been mixed. Ades (1974) used both CV and VC syllables and found no evidence for adaptation effects either in the direction of the spectral overlap between adaptor and test series or the common phonetic identity. Pisoni (Note 2; also Pisoni & Tash, 1975) reported that adaptation with a VC produced an effect on a CV test series that followed spectral overlap and was opposite the phonetic identity of the adaptor. Sawusch (1977b) found results similar to Ades (1974) in that VC adaptors failed to affect a CV test series, although nonspeech, VC-like adaptors did produce adaptation effects on the CV test items with similar spectral properties. Finally, Wolf (1978) found some weak evidence in support of adaptation in the direction of the phonetic identity of the adaptor (Experiment I), although, as she notes, subsequent results (Experiment II) suggested that these adaptation effects were probably due to spectral similarities between the vocalic portion of the adaptor and the consonantal portion of the test items.

One factor which complicates the interpretations of the above studies is that in each instance, the consonant undergoing adaptation occurred in a different syllable position, with respect to the vowel, in the adaptor than in the test items. For example, in Ades' study, [aed] served as an adaptor for a [bae] - [dae] continuum. If, as Cooper (1975) has suggested, consonants following vowels are processed differently than those preceding vowels, then the test conditions employed in these studies may not provide a satisfactory context for evaluating the relative contributions of the spectral characteristics and phonetic identity of the adaptor. Instead, what is required is an experimental setting in which the consonant undergoing adaptation occurs in the same syllable position, relative to the vowel, in both the adaptor and the test items. Experiment 1 was designed to provide just such a test situation.

Experiment 1

In English, the labial stop consonant which follows an [s] in an initial consonantal cluster is identified as the voiceless segment [p]. However, in an early perceptual experiment, it was shown that if the initial [s] was spliced out of the stop-fricative cluster, then English speakers perceive the remaining stop segment as the voiced labial [b] (Lotz, Abramson, Gerstman, Ingemann & Nemser, 1960). More recently, Klatt (1975; see also Davidsen-Nielsen, 1974) has measured the voice-onset times of labial stop segments occurring within word-initial consonant clusters. His measurements showed that the labial stop in an [sp] cluster had an average VOT of +12 msec; one which coincided much more closely to the values found for syllable initial [b] (+11 msec) than for syllable initial [p^h] (+47 msec). Nevertheless, despite these indications that the labial stop in [sp] clusters is spectrally similar to [b], the fact remains that these segments are assigned to the phonetic category [p], given the phonological context in which they occur in English.

One possible way of assessing the relative contributions of spectral characteristics and phonetic identity to an adapting stimulus would be to employ [spa] as an adaptor and test for changes in the identification of a [ba] - [p^ha] continuum. The use of such a test would help to distinguish among the three hypotheses outlined above because different predictions follow from each one. According to the first position, which we will refer to as the spectral-overlap hypothesis, a [spa] adaptor should have an adaptation effect that follows the spectral overlap with the test series. That is, it would be expected to behave in a manner similar to a [ba] adaptor. The predictions according to the second position, the two level model, depend upon the nature of the second level of processing. If the second level is assumed to be phonetic, as Cooper (1975) suggests, then the spectral overlap and the perceived phonetic quality of the [spa] adaptor should act in opposite directions and cancel each other out. The net result should be no observable adaptation effect. If the second level is assumed to be abstract auditory, one which integrates information over a set of acoustic cues (Sawusch, 1977a), then the [spa] adaptor, which carries the same spectral and pattern information as the [ba] end of the test series, should have an adapting effect identical to the [ba]. Thus, the two level model of Sawusch (1977a) makes basically the same predictions as the spectral overlap position. Finally, the third hypothesis, based on response contrast, predicts that the [spa] adaptor's effect on the test series should follow its perceived quality since only phonetic decision rules or the subjects decision criteria are modified. To the extent that this stimulus is perceived (labeled) as [p], its effect on the [ba] - [p^ha] test series should be similar to the influence of the [p^ha] endpoint. That is, the effects should be similar to those reported by Diehl et al. (1978).

In addition to the [spa] stimulus (hereafter referred to as [sba]), several other adapting stimuli were also employed in this experiment. First, in order to obtain some measure of the degree of adaptation induced with [sba], all subjects participated in test sessions in which the [ba] and [p^ha] endpoints served as adaptors. Second, all subjects were tested for one session using a [sp^{h}a] adaptor; a syllable constructed by placing the fricative [s] plus a silent period in front of the [p^{h}a] endpoint stimulus. This adaptation condition was included to determine whether a failure to find adaptation with [sba] could be attributed to the presence of conflicting spectral and phonetic cues or to some form of interference induced by the presence of the initial [s]. Note that all three hypotheses predict that any adaptation effects with [sp^{h}a] should result in a shift of the phonetic category boundary toward [p^{h}a]. Finally, all subjects underwent an adaptation session using a [sla] adaptor; a syllable produced by directly concatenating an [s] to the beginning of the [ba] endpoint with only a 5 msec closure period. This last condition was included in order to assess whether the spectral overlap between the adaptor and the [ba] endpoint would be sufficient to induce a boundary shift even under circumstances in which no stop was perceived to be present in the adaptor.

Method

Subjects. The subjects were nine undergraduates at Dalhousie University. All were native speakers of English and reported no history of either speech or hearing disorder. Each subject participated in five one-hour sessions on separate days and was paid $17.50 (Canadian) upon completion of the experiment.

Stimuli. The stimuli were eleven synthetic speech sounds prepared on the PDP 11/05 computer at the Speech Perception Laboratory at Indiana University. All stimuli were generated on the cascade-parallel synthesizer originally designed by Klatt (Note 3; see Kewley-Port, Note 4). Eight of the stimuli were labial stop consonant-vowel syllables which ranged perceptually from a 0 msec VOT [ba] to a +70 msec VOT [p^{h}a] in 10 msec VOT steps. The values used for the synthesis were chosen from measurements of natural speech produced by a male talker and are identical to those used by Pisoni, Aslin, Perey & Hennessy (Note 5). The stimuli consisted of a 255 msec steady-state pattern with formant values appropriate for the vowel [a] (F1 = 700 Hz, BW1 = 90 Hz; F2 = 1200 Hz, BW2 = 90 Hz; F3 = 2600 Hz, BW3 = 130; F4 = 3300 Hz, BW4 = 400; and F5 = 3700, BW5 = 500). The formant transitions in front of the vowel were 40 msec in duration and had starting frequencies of 438 Hz (F1), 1025 Hz (F2) and 2425 Hz (F3). Each stimulus also contained a 10 msec release burst generated by using the parallel branch of the synthesizer in conjunction with a turbulent noise source (AF). The variation in VOT in this series was produced by replacing the vocalic excitation (AV) with a turbulent noise source (AH) in the cascade branch and simultaneously widening the bandwidth of F1 to 300 Hz for the duration of the noise. The stimuli had a pitch contour which rose from 120 Hz to 125 Hz during the first 50 msec and then fell linearly to 100 Hz at syllable offset.

Three additional consonant-consonant-vowel (CCV) syllables were also generated by prefixing an [s] to either the [ba] endpoint or the [p^{h}a] endpoint of the VOT series. For two of these stimuli, the [s] was prefixed to the 0 msec VOT [ba]. The only difference between these two syllables was in the duration of the silent interval between the offset of the [s] and the onset of the [ba]. For the [sba] stimulus,[1] the duration of the interval was 75 msec, while for the [sla] stimulus, the interval duration was 5 msec. In all other respects, these two stimuli were identical. For the third CCV, [sp^{h}a], the [s] was added to the initial portion of the +70 msec VOT [p^{h}a]. A silent interval of 75 msec between the [s] offset and the [p^{h}a] onset was used in this syllable (the same interval as used in the [sba] stimulus). In all three syllables, the [s] was 145 msec in duration and was generated by exciting the parallel branch of the synthesizer with a turbulent noise source (AF). Only F5 and F6 were used in generating the [s] (the amplitudes of F1 through F4 were set to zero). The center frequencies and bandwidths of F5 and F6 were 3900 Hz, 1000 Hz, 4900 Hz and 1000 Hz respectively.

Twelve test tapes were generated using an audio tape making program at Indiana University. The stimuli were converted to analogue form in real-time via a 12-bit digital-to-analogue converter, low-pass filtered at 4.8 kHz and recorded on a Crown model 822 tape recorder. Two of the tapes contained different randomized identification sequences of the eight syllable VOT test series. There were ten occurrences of each syllable for a total of 80 trials per tape. A four sec response interval separated successive syllables on each tape. The remaining ten tapes contained adaptation sequences; two tapes for each of the following adaptor syllables: 0 msec VOT [ba], +70 msec VOT [p^{h}a], [sba], [sp^{h}a] and [sla]. Each tape consisted of an initial 2 min adaptation period during which time the adapting stimulus was presented repeatedly with a 300 msec interadaptor interval. After a four sec pause, which cued listeners that the identification test trials were about to start, all eight stimuli from the VOT series were presented once, in random order, for identification. At the conclusion of a block of eight test trials, there was a four sec pause followed by 75 more repetitions of the adaptor. Another block of eight identification trials was then presented in a different random order. The alternation of adaptor repetitions and test trials was repeated ten times per tape.

Procedure. Each subject was tested individually in a small quiet room. Practice and test tapes were played on a Teac 3340S tape recorder equipped with Koss PRO 4AAA headphones. The volume was adjusted with a sound level meter (General Radio model 1565-A) so that the stimuli were played at a level of approximately 72 dB(A) SPL.

On the first day of testing, all subjects were presented with a practice identification sequence consisting of 24 trials (three repetitions of each of the eight stimuli). Subjects were asked to make two responses to each syllable. First, they were asked to identify each sound as either [ba] or [p^ha] and write their response ("B" or "P") on an answer sheet. They were also asked to indicate how confident they were that they had identified the syllable correctly using a 4-point scale. A copy of the scale was in view for the subjects throughout testing (1 - positive; 2 - probable; 3 - possible; 4 - guess). In order to ensure that subjects were comfortable with the response rating procedure, the practice tape was played twice. Upon completion of practice, testing began. On Day 1, subjects were presented with two identification tests. Following this, adaptation testing was initiated with one of the five adaptors. Subjects were told that they would hear a syllable repeated a number of times, followed by a pause and then eight syllables that they were to identify as "B" or "P" and give their confidence rating. On Days 2 through 5, the testing procedure was similar except that the practice tapes were omitted and only one baseline identification tape was played. On any particular day, subjects listened to both of the adaptation tapes for one adaptor. For each subject, the order in which the five adaptation conditions were presented was randomized across days. By the end of the experiment, each subject had provided at least 20 responses to each of the test stimuli under both identification and each of the five adaptation conditions.

Results

One subject was dropped from the experiment after the first day for failing to follow instructions (he talked while listening to the tapes). The baseline identification performance for each of the remaining eight subjects was examined for the second block of identification trials on Day 1 and those of Days 2 through 5. A repeated measures one-way ANOVA revealed no significant differences across days (F(4,28) = 1.0). Consequently, the data from these blocks were collapsed to provide a baseline measure of identification with 100 observations per data point per subject. Examination of the rating data indicated that subjects tended to use only the extreme positive (1) rating. Therefore, the rating data were omitted from further analysis, and only the "B" and "P" identification responses were employed. The locus of each subject's [ba] - [p^ha] category boundary was calculated using the mean from a standard ogive measure (see Engen, 1971). The same ogive measure was also applied to determine the location of each subject's boundary after adaptation with each of the five adaptors. Table 1 displays each subject's unadapted baseline category boundary, along with the amount and direction of change in the boundary location after exposure to each of the adaptors.

Insert Table 1 about here

Correlated t-tests were used to determine whether or not a significant shift occurred in the boundary locus after adaptation. As expected, the [ba] and [p^ha] adaptors had opposite effects. Adaptation with [ba] resulted in a significant shift in the category boundary toward [ba] (*t*(7) = -3.32, *p* < .01, one-tailed) while [p^ha] adaptation led to a reliable shift toward [p^ha] (*t*(7) = 2.189, *p* < .05, one-tailed). These results concur with previous findings using similar stimuli (Eimas & Corbit, 1973; Tartter & Eimas, 1975).

Of more importance, however, are the results for the three consonant-consonant-vowel adaptors. The [sp^ha] adaptor produced an effect similar to that of the [p^ha] adaptor (*t*(7) = 4.82, *p* < .002, two-tailed). This result was not surprising since both the phonetic label and the spectral cues mark the stop in this adaptor as [p]. The [sba] adaptor, however, has a phonetic label which conflicts with its spectral cues. Thus, it is interesting to note that adaptation with [sba] resulted in a significant shift in the category boundary toward [ba] (*t*(7) = -3.39, *p* < .02, two-tailed). In other words, in this case adaptation followed the spectral overlap between the adaptor and the test series and not the perceived quality of the adaptor. Finally, the [sla] adaptor, like the [sba] adaptor, contained spectral cues appropriate for a [b]. Nevertheless, none of our subjects ever reported hearing a stop in this adaptor. Most described the adaptor as "SLA", or occasionally "SA". Adaptation with this item produced a small, but nonsignificant, shift in the category boundary toward [ba] (*t*(7) = -1.72, .1 < *p* < .2, two-tailed). Thus, in this instance, it would appear that the spectral overlap between the adaptor and the test series was not sufficient to produce a consistent shift in the category boundary.

Discussion

The major finding of the present experiment is that adaptation with [sba] produces a shift of the phonetic category boundary towards [ba]. This result is directly in line with the predictions of the spectral overlap hypothesis, and in contradiction to the response contrast hypothesis. By the latter hypothesis, adaptation with [sba] should have shifted the phonetic boundary toward [p^ha] since the perceived phonetic identity of this fricative-labial stop cluster adaptor was [p]. The present results also indicate that if there are two levels of processing that are affected by adaptation, neither of these levels is a phonetic level of processing. If two levels of processing were operating and one of these levels was phonetic in nature, then we should have found little or no adaptation effect for the [sba] adaptor since adaptation at auditory and phonetic levels would have been in opposite directions and cancelled each other. Since the effect of the [sba] adaptor was nearly identical to that of the [ba] and these two adaptors were identical spectrally (for their stop-vowel portions) but opposite in their phonetic identity, it seems safe to conclude that no phonetic processes were affected by adaptation in this experiment.

There are two possible objections which could be raised at this juncture with respect to the present account. The first is that although spectral overlap was sufficient to induce adaptation for the [sba] adaptor, it was not sufficient for the [sla] adaptor. In this connection, it should be noted, that although not significant, the boundary shifts after adaptation with [sla] were in the correct direction according to the spectral overlap hypothesis (i.e. towards [ba]). Moreover, it would appear that the failure to achieve a significant effect of adaptation in this case can be traced to the performance of one subject who displayed a boundary shift towards [p^ha]. As a check on this conjecture, an additional eleven subjects were run in the adaptation procedure with the [sla] stimulus as the only adaptor used.[2] Of the eleven subjects, one showed a shift toward [p^ha], one showed no shift and nine showed small shifts in their [ba] - [p^ha] category boundary toward [ba]. The mean shift of .11 stimulus units (1.1 msec VOT) was significant ($\underline{t}(10) = -3.52$, $\underline{p} < .01$ two-tailed). The magnitude of the effect for these eleven additional subjects is nearly identical to that found for the [sla] adaptor in the original experiment (see Table 1). Furthermore, the effect of the [sla] adaptor was significantly smaller than both the [ba] and the [sba] adaptation effects ($\underline{t}(17) = 2.50$, $\underline{p} < .05$ and $\underline{t}(17) = 2.42$, $\underline{p} < .05$ respectively, two-tailed). Thus, the [sla] adaptor does have a consistent, though small, adapting effect on the test series and the direction of this effect follows the spectral overlap between the adaptor and the test series.

The second objection to our interpretation of the present experiment is that, contrary to our expectations, subjects may not have perceived the [sba] stimulus as "SPA" but rather as some "S"-like noise plus "BA". If such were the case, the present experiment would have failed to distinguish between the three hypotheses since the phonetic identity would have coincided with the spectral characteristics of the labial stop in the adaptor. However, this explanation of the results of experiment seems unlikely since all subjects reported hearing the [sba] adaptor as "SPA". Nevertheless, in order to better assess the way in which subjects were perceiving the adapting stimuli used in the present experiment, a second experiment was conducted using a procedure similar to that employed by Diehl et al. (1978).

<u>Experiment 2</u>

In order to distinguish whether the spectral composition of an adapting stimulus or its perceived quality is responsible for the adapting effects found in Experiment 1, it is necessary to demonstrate either that the [sba] was, in fact, perceived as "SPA", or that the perceived quality of the [sba] is irrelevant to and independent of its effects as an adaptor. Unfortunately, Experiment 1 did not completely satisfy either of these conditions since it is possible that the subjects did perceive the [sba] as more "B"-like than "P"-like. If this were the case, then the perceived quality of the adaptor could be controlling the adaptation effects found. This line of reasoning is consistent with a recent set of experiments reported by Diehl and his coworkers (Diehl et al., 1978; in press). As outlined above, in these experiments, stimuli were presented in pairs. In some of the pairs, a near boundary stimulus was paired with a good exemplar of a phonetic category. Their typical results indicated that a contrast effect was occurring. Identification of the near boundary stimulus consistently fell into response categories other than that applied to the good exemplar (context item). Diehl et al. (1978; in press) have termed this effect "response contrast" and have proposed it as an alternative explanation for selective adaptation results. The main thrust of their argument is that both selective adaptation and response contrast experiments can be seen as versions of the same psychophysical paradigm; anchoring. The chief difference between the two consists in the number of times that the anchoring or reference stimulus is repeated. They question the validity of earlier studies which purport to show that adaptation effects cannot be accounted for by a response contrast hypothesis (e.g. Cooper, 1974; Eimas & Corbit, 1973; Diehl, 1976). In support of their position, they cite Elman's (1979) finding that virtually all of the changes in subjects' performance after adaptation could be accounted for by a change in decision criterion, as opposed to a change in sensitivity.

The procedure used by Diehl et al. would seem to offer a method of resolving whether the effect of our [sba] adaptor was due to spectral overlap with the test series or its perceived quality. From the results of Experiment 1, the 30 msec VOT test stimulus appears to be close to the baseline category boundary (see Table 1, the 30 msec VOT item is Stimulus 4). If this stimulus is presented to subjects as part of a pair with the various adaptors from Experiment 1, then it should be possible to determine whether it is the perceived quality of the adaptors that influences adaptation of the test items. Specifically, if the adapting effect of the [sba] adaptor was due to its being perceived as "B", then the influence of the [sba] in the Diehl et al. procedure should be identical to the influence of the [ba] in the same procedure. That is, both [sba] and [ba] should cause the 30 msec VOT test item to sound more "P"-like. Alternatively, if the [sba] is perceived by the subjects as "P" then opposite effects should be found for the [ba] and [sba] contexts on the test item. The [ba] should induce more "P" responses while the [sba] induces more "B" responses. This result would indicate that the influence of the [sba] adaptor in Experiment 1 was clearly due to its spectral structure and was entirely unaffected by the perceived quality of the adaptor. Finally, it is also possible that the [sba] context item could have little or no influence on the test item. This could arise from either the dissimilarity of the [sba] to the test item, due to the initial [s], or because the perceived quality of the [sba] is, in fact, ambiguous between "B" and "P". In either case, the absence of an effect of the [sba] on the test item would rule out perceived quality as an explanation of the adaptation effects of [sba] since the [sba] stimulus did have an adapting effect identical to that of a [ba] in Experiment 1. Thus, the use of the [sba] stimulus in the procedure described by Diehl et al. (1978; in press) allows a strong test of their claim that part, if not all, of the category boundary shifts found in selective adaptation experiments are a result of response contrast (perceived similarity between adaptor and test item).

Method

Subjects. The subjects were fifteen undergraduate and graduate students at the State University of New York at Buffalo. All were right-handed, native speakers of English with no reported history of either speech or hearing disorder. All subjects participated for one hour and were paid $3 (U.S.).

Stimuli. The stimuli used in this experiment consisted of the 0, 10, 30, 50, and 70 msec VOT items from the [ba] - [p^{h}a] series used in Experiment 1 and the [sba], [sp^{h}a], and [sla] stimuli that were used as adaptors in Experiment 1. The stimuli were generated on a PDP 11/34 computer in the Speech Perception Laboratory at SUNY/Buffalo using the cascade-parallel synthesizer developed by Klatt (Note 3; see Kewley-Port, Note 4).

Procedure. The subjects were run in small groups of three or four at a time. The format for all subjects was the same. The stimuli were presented to the subjects in real-time by a PDP-11/34 computer in the Speech Perception Laboratory at SUNY/Buffalo. The stimuli were converted to analogue form via a 12-bit D/A converter and presented binaurally at an intensity of 75 dB SPL via TDH-39 matched and calibrated headphones. All stimuli were presented to the subjects in pairs. One half of the pairs were test pairs and the other half were filler pairs. In a test pair, the 30 msec VOT test item was presented along with one of five other context stimuli: (1) the 0 msec VOT [ba], (2) the 70 msec VOT [p^{h}a], (3) the [sba], (4) the [sp^{h}a], or (5) the [sla]. In the filler pairs, the 10 msec VOT [ba] and the 50 msec VOT [p^{h}a] were always presented together. In all pairs, the two stimuli were separated by 500 msec (ISI). The order of stimuli within each pair was counterbalanced so that each order occurred on half of the trials. The subjects were asked to listen to both of the stimulus items in a pair. After the second item, they were to identify both of the stimuli in their order of occurrence. Subjects were asked to identify each stimulus as either "BA" or "PA" and enter this response by pushing the appropriate button on a response box in front of them. Subjects were informed that some of the stimuli would have an "S" sound at the beginning but that each stimulus should be identified as either "BA" or "PA", even if they had to guess.

The order of presentation of test pairs was random. Each test pair was always followed by a filler pair. All subjects listened to four blocks of 80 pairs (half of which were test pairs and half filler pairs). Within a block of trials, each test pair occurred eight times, four times with the test stimulus followed by the context stimulus and four times with the reverse order. The interval between pairs was 5 sec. Subjects were given a short break between blocks. By the end of the experiment, each subject had provided a total of 32 responses to each item in the test pairs.

Results

The number of "BA" responses to the test item when paired with each of the five context items was tabulated for each subject. These totals appear in Table 2. The data for two of the fifteen subjects were eliminated because these subjects identified the 50 msec VOT filler item as a "BA" more than 20% of the time, indicating that the 30 msec VOT test item was not close to the category boundary for these subjects. The [ba] context led to significantly fewer "BA" responses on the test item than the [p^{h}a] context (*t*(12) = 3.17, *p* < .01 for the mean difference of 6.92 responses). These results replicate those of Diehl et al. (1978). The [sp^{h}a] context produced results nearly identical to the [p^{h}a] context. Significantly more "BA" responses were found for the test item when paired with the [sp^{h}a] context than the [ba] context (*t*(12) = 2.41, *p* < .05 for a mean difference of 5.92 responses). Moreover, the [sp^{h}a] and [p^{h}a] contexts did not differ significantly from each other (*t*(12) = 1.45, *p* > .10 for a mean difference of 1.00 responses). On the other hand, the effect of the [sba] context was significantly different from both the [ba] context and the [p^{h}a] context. More "BA" responses on the test item were found when it was paired with [sba] than when it was paired with [ba] (*t*(12) = 2.35, *p* < .05 for a mean difference of 3.54 responses) and fewer "BA" test item responses were found with the [sba] context than with the [p^{h}a] context (*t*(12) = 3.42, *p* < .01 for a mean difference of 3.38 responses). The [sla] context produced results very similar to the [sba] context. The number of "BA" responses to the test item was significantly higher when paired with the [sla] context than when paired with the [ba] context (*t* (12) = 2.32, *p* < .05 for a mean difference of 3.69 responses) and significantly lower for the [sla] context compared to the [p^{h}a] context (*t*(12) = 3.36, *p* < .01 for the mean difference of 3.23 responses).

Insert Tables 2 and 3 about here

The number of "BA" and "PA" responses to the [sba], [sp^{h}a] and [sla] contexts was also tabulated and appear in Table 3. Each of the thirteen subjects showed more "PA" responses than "BA" responses to the [sp^{h}a] context. Of the 416 total presentations of this context item, only two received "BA" responses. The [sba] context, by comparison, did not receive significantly more "BA" responses than "PA" responses. Of the thirteen subjects, eight showed more "BA" responses while five showed more "PA" responses (p > .20 using a sign test). Thus, the [sp^{h}a] context represented a good "PA" to the subjects while the [sba] context seems to have been ambiguous between a "BA" and a "PA". The [sla] context yielded more "BA" responses than "PA" responses for ten of the subjects while two showed more "PA" responses. However, during the debriefing session after the experiment, subjects reported that this stimulus sounded more like "SLA" or "SA". Thus, to the extent that the [sla] context was not consistently heard as either a "BA" or a "PA" it was also ambiguous.

Discussion

The influence of the [ba], [p^{h}a] and [sp^{h}a] contexts on the test item was one of contrast. The test item was often identified as belonging to a category other than that of the context item it was paired with. These results are similar to the adaptation results of Experiment 1 since the category boundary shifts found there were also in a contrastive direction. Therefore, the present results replicate those of Diehl et al. (1978) and extend them to include the [sp^{h}a] context. However, the [sba] context did not produce a contrast effect since identification of the test item, when paired with [sba], was intermediate between the results with [ba] and [p^{h}a] contexts. Thus, the [sba] seems to have had little or no influence upon the identification of the target item. A similar pattern was found for the [sla] context.

These results demonstrate that the perceived quality of the context items determines their influence upon the test item in the Diehl et al. (1978) procedure. The [sp^{h}a] context stimulus, which was identified by subjects as "P", had virtually the same effect on the test item (causing more "B" responses) as the [p^{h}a] context stimulus. However, both the [sba], which was identified inconsistently as either "B" or "P", and the [sla], which did not have a stop-like quality at all, had influences on the test item that were intermediate between the [ba] and [p^{h}a] contexts. Thus, the perceived quality of the context item seems to consistently predict its influence upon the test item in this experimental procedure.

If the spectral overlap between context and test item were responsible for the contrast effects observed in this procedure, then the [ba] and [sba] contexts should have produced identical results, even though the [sba] was not consistently identified as "B". Similarly, if the results of this procedure were dependent on the identity of the context item according to the phonological constraints of English, then the [sba], [p^{h}a], and [sp^{h}a] contexts should all have produced equivalent results. Since neither of these outcomes was found, it seems reasonable to conclude that the perceived quality of the [sba] context (intermediate between "B" and "P") was responsible for its influence upon the test item. A similar conclusion can be drawn for the effect of the [sla] context. Thus, the context effects reported by Diehl et al. (1978; in press) and in the present experiment using the same procedure seem to operate at a level of processing based on the perceptual similarity between the context and the test item. Spectral overlap between the context and test items does not appear to be an important factor in these results. This dissociation of the effects from the two procedures has important consequences for the interpretation of selective adaptation results using speech stimuli.

General Discussion

The results of Experiment 2 demonstrate that the perceived quality of the context item is responsible for its influence on the test item paired with it. The response contrast explanation of Diehl et al. (1978; in press) seems to be consistent with this description. By comparison, it was the spectral overlap between adaptor and test item that determined the adaptation effects found in Experiment 1. The direction and magnitude of the shifts in the phonetic category boundary found in Experiment 1 with the [sba] and [sla] adaptors seem to be independent of, and in the case of [sba], opposite to that expected on the basis of the response contrast explanation. Consequently, the results found using the Diehl et al. procedure appear to be unrelated to and independent of the effects found in selective adaptation experiments, even though there is some superficial similarity across these two procedures. Taken together, the results of Experiments 1 and 2 demonstrate that response contrast is not involved in the selective adaptation results found for voicing in stop consonants.

This study also poses serious problems for the view that there is a phonetic level of processing which is affected by selective adaptation along a voicing continuum. If a phonetic level were present, then the

[sba] adaptor should have had little or no net effect upon the [ba] - [p^ha] test series. However, the [sba] and [ba] adaptors produced virtually identical effects, despite their opposite phonetic identities. Thus, if two levels of processing are involved in selective adaptation effects for voicing information, neither of these levels is phonetic. Sawusch (1977a) has offered a similar account regarding place of articulation in stop consonants.

Having ruled out both a response contrast explanation and the involvement of phonetic processes in selective adaptation results, we are left with two alternatives. One is the operation of auditory feature detectors which are sensitive to the spectral overlap between adaptor and test item (Ades, 1976; Bailey, 1975). The second is a two-level model involving an abstract auditory level of processing, which is sensitive to configurations of cues in the acoustic signal, in addition to local auditory feature detectors. The present experiments cannot distinguish between these two models. However, they do provide some further specification of the characteristics of selective adaptation to voicing information in stop consonants. As the results of Experiment 1 clearly demonstrated, adaptation follows the spectral overlap between adaptor and test series. However, while the [sba] and [ba] adaptors produced virtually identical results, the [sla] adaptor produced a much smaller change in the identification of the test series. This implies that the period of silence, corresponding to articulatory closure in stop production, that is present in the [sba] (and [ba]), but not in the [sla], was registered by the perceptual mechanism(s) being adapted. Thus, adaptation was not operating entirely at a level which simply registered the onset relationship between the first and second formants, a sufficient cue for the voiced-voiceless distinction in initial

English stops (see Liberman, Delattre & Cooper, 1958). Rather, adaptation seems to be based on the configurational properties of the stimulus as a whole.

Taken together with the results of other studies (Ganong, 1978; Pisoni & Tash, 1975; Sawusch, 1977a, b) the present results demonstrate that adaptation follows the auditory pattern information within the speech signal. In the present study, adaptation followed the spectral overlap between the adaptor and the test series. When the phonetic identity of the adaptor conflicted with its spectral characteristics, the spectral characteristics goverened the direction and magnitude of the adaptation effects that were found. Our results also demonstrated the effects found with a paired comparison procedure (Diehl et al., 1978) do not always mirror selective adaptation results. Rather, the contrast effects found in the paired comparison procedure seem to arise from a similarity in the perceived quality of the items paired together. Selective adaptation, however, seems to operate at an earlier, auditory level of processing, prior to the determination of the phonetic identity of the stimulus item.

Reference Notes

1. Ganong, W. F. An experiment on "phonetic adaptation". RLE Progress Report, MIT, Cambridge, Massachusetts, 1975, 116, 206-210.

2. Pisoni, D. B. Stages of processing in speech perception: Feature analysis. Paper presented at the Eighth International Congress of Phonetic Sciences, August, 1975, Leeds, England.

3. Klatt, D. H. A cascade/parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis. Paper presented at the 93rd meeting of the Acoustical Society of America, June, 1977, University Park, Pennsylvania.

4. Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. Research on Speech Perception, 1978, Progress Report 4, Indiana University.

5. Pisoni, D. B., Aslin, R. N., Perey, A. J. & Hennessy, B. L. Identification and discrimination of a new linguistic contrast. Research on Speech Perception, 1978, Progress Report 4, Indiana University.

References

Ades, A. E. How phonetic is selective adaptation? Experiments on syllable position and vowel environment. *Perception & Psychophysics*, 1974, *16*, 61-67.

Ades, A. E. Adapting the property detectors for speech perception. In R. J. Wales & E. Walker (Eds.), *New approaches to language mechanisms*. Amsterdam: North-Holland, 1976.

Bailey, P. J. Perceptual adaptation in speech: Some properties of detectors for acoustical cues to phonetic distinctions. Unpublished doctoral dissertation, 1975, University of Cambridge, Cambridge, England.

Cooper, W. E. Selective adaptation for acoustic cues of voicing in initial stops. *Journal of Phonetics*, 1974, *2*, 303-313.

Cooper, W. E. Selective adaptation to speech. In F. Restle, R. M. Shiffrin, N. J. Castellen, H. Lindman and D. B. Pisoni (Eds.), *Cognitive theory: Vol 1*. Hillsdale, N.J.: Earlbaum, 1975.

Cooper, W. E. & Blumstein, S. A "labial" feature analyzer in speech perception. *Perception & Psychophysics*, 1974, *15*, 591-600.

Davidsen-Nielsen, N. Syllabifaction in English words with medial sp, st, sk. *Journal of Phonetics*, 1974, *2*, 15-45.

Diehl, R. L. Feature analyzers for the phonetic dimension *stop vs continuant*. *Perception & Psychophysics*, 1976, *19*, 267-272.

Diehl, R. L., Elman, J. L. & McCusker, S. B. Contrast effects in stop consonant identification. *Journal of Experimental Psychology: Human Perception and Performance*, 1978, *4*, 599-609.

Diehl, R. L., Lang, M. & Parker, E. M. A further parallel between selective adaptation and response contrast. *Journal of Experimental Psychology: Human Perception and Performance*, in press.

Eimas, P. D., Cooper, W. E. & Corbit, J. D. Some properties of linguistic feature detectors. *Perception & Psychophysics*, 1973, *13*, 247-252.

Eimas, P. D. & Corbit, J. D. Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, 1973, *4*, 99-109.

Eimas, P. D. & Miller, J. L. Effects of selective adaptation of speech and visual patterns: Evidence for feature detectors. In H. L. Pick & R. D. Walk (Eds.), *Perception and Experience*. New York: Plenum, 1978.

Elman, J. L. Perceptual origins of the phoneme boundary effect and selective adaptation to speech: A signal detection theory analysis. *Journal of the Acoustical Society of America*, 1979, *65*, 190-207.

Engen, T. Psychophysics I. Discrimination and detection. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth and Schlosberg's Experimental Psychology*. New York: Holt, Rinehart and Winston, 1971.

Ganong, W. F. The selective adaptation effects of burst-cued stops. *Perception & Psychophysics*, 1978, *24*, 71-83.

Klatt, D. H. Voice onset time, friction and aspiration in word initial consonant clusters. *Journal of Speech and Hearing Research*, 1975, *18*, 686-706.

Liberman, A. M., Delattre, P. C. & Cooper, F. S. Some cues for the distinction between voiced and voiceless stops in initial position. *Language and Speech*, 1958, *1*, 153-167.

Lotz, J., Abramson, A. S., Gerstman, L. J., Ingemann, F. & Nemser, W. J. The perception of English stops by speakers of English, Spanish, Hungarian, and Thai: A tape cutting experiment. *Language and Speech*, 1960, *3*, 71-77.

Miller, J. L., Eimas, P. D. & Root, J. Properties of feature detectors for place of articualtion. *Journal of the Acoustical Society of*

America, 1977, 61, S48 (A).

Pisoni, D. B. & Tash, J. B. Auditory property detectors and processing place features in stop consonants. Perception & Psychophysics, 1975, 18, 401-408.

Remez, R. E. Adaptation of the category boundary between speech and nonspeech: A case against feature detectors. Cognitive Psychology, 1979, 11, 38-57.

Remez, R. E., Cutting, J. E. & Studdert-Kennedy, M. Acoustic similarity or phonetic identity: A cross adaptation study employing song and string. Perception & Psychophysics, in press.

Sawusch, J. R. Peripheral and central processes in selective adaptation of place of articualtion in stop consonants. Journal of the Acoustical Society of America, 1977, 62, 738-750 (a).

Sawusch, J. R. Processing place information in stop consonants. Perception & Psychophysics, 1977, 22, 417-426 (b).

Simon, H. J. & Studdert-Kennedy, M. Selective anchoring and adaptation of phonetic and nonphonetic continua. Journal of the Acoustical Society of America, 1978, 64, 1338-1357.

Tartter, V. C. & Eimas, P. D. The role of auditory and phonetic feature detectors in the perception of speech. Perception & Psychophysics, 1975, 18, 293-298.

Wolf, C. G. Perceptual invariance for stop consonants in different positions. Perception & Psychophysics, 1978, 24, 315-326.

Footnotes

This work was supported by NIMH grant MH 31468-01 and NSF grant BNS 7817068 to SUNY/Buffalo, an N.S.E.R.C. grant (A-0282) to the second author and NINCDS grant NS-12179-04 to Indiana University at Bloomington. The authors would like to thank Dr. David B. Pisoni for making the facilities of the Speech Perception Laboratory at Indiana University available for preparing the stimuli and tapes used in Experiment 1 as well as for his helpful comments on an earlier version of this manuscript. Special thanks are also extended to G. Eleanor Stewart for her assistance in running subjects in Experiment 1 and to Eileen C. Schwab for running subjects in Experiment 2. Reprint requests may be sent to either author.

[1] We have chosen to refer to this stimulus as [sba] strictly for mnemonic purposes, i.e. to stress its spectral similarity to [ba]. The correct phonetic designation is actually [spa], but, for that matter, the proper phonetic designations for the English "ba" and "pa" are [pa] and [p^ha] respectively, even though they are typically represented as [ba] and [pa].

[2] The subjects were eleven undergraduates at SUNY/Buffalo who participated to fulfill a course requirement. The procedure was basically the same as described above. Subjects were run in small groups of one to four at a time. The presentation of stimuli, timing of intervals and collection of subject responses was all done under computer control. All subjects listened to twenty occurrances of each of the eight test stimuli in random order (in two blocks of 80 trials), followed by adaptation trials as described previously. By the end of testing, each subject had provided 20 baseline and 20 [sla] adapted identification responses to each of the test

stimuli. Category boundaries for both the baseline and [sla] adapted conditions were calculated for each subject as previously described.

Table 1

Individual Subject's Boundaries for the VOT Series and the Resultant Change in Locus Due to Each Adaptor.*

Subject	Unadapted	Change in Boundary After Adaptation				
	VOT Boundary	[ba]	[p^ha]	[sba]	[sp^ha]	[sla]
1	3.50	.00	.47	.00	.55	.00
2	4.18	-.39	.23	-.98	-.01	-.13
3	3.78	-.28	.13	-.28	.13	-.28
4	3.50	.00	.00	-.29	.29	-.29
5	3.36	-.72	.14	-.27	.50	-.33
6	3.83	-.82	.10	-.38	.55	-.33
7	3.83	-.33	.19	-.33	.37	.00
8	4.08	-.25	-.15	-.10	.37	.29
Mean	3.76	-.35	.14	-.34	.34	-.13

* Note that a negative sign indicates a shift in the category boundary toward the [ba] end of the series while a positive shift is toward the [p^ha] end of the series.

Table 2

Number of "BA" Responses to the Test Stimulus (30 msec VOT)

When Paired With Each of Five Context Stimuli

	Context Stimulus				
Subject	[ba]	[p^{h}a]	[sba]	[sp^{h}a]	[sla]
1	6	22	10	23	14
2	3	31	22	32	22
3	28	30	28	26	29
4	21	27	22	22	22
5	31	32	31	30	28
6	30	31	30	31	30
7	26	30	28	31	30
8	2	9	7	13	11
9	5	12	7	9	6
10	19	20	20	19	20
11	20	20	20	20	20
12	15	20	17	17	16
13	16	28	26	26	22
Mean	17.08	24.00	20.62	23.00	20.77

Table 3

Number of "BA" and "PA" Responses to the Context Stimuli [sba], [sp^{h}a], and [sla]

	[sba]		[sp^{h}a]		[sla]	
Subject	"BA"	"PA"	"BA"	"PA"	"BA"	"PA"
1	32	0	0	32	32	0
2	22	10	0	32	19	13
3	32	0	0	32	32	0
4	20	12	1	31	17	15
5	26	6	0	32	32	0
6	21	11	0	32	25	7
7	5	27	0	32	16	16
8	10	22	0	32	32	0
9	1	31	0	32	24	8
10	24	8	0	32	12	20
11	32	0	0	32	6	26
12	6	26	1	31	17	15
13	10	22	0	32	27	5

II. SHORT REPORTS AND WORK-IN-PROGRESS

Some Remarks on the Perception of Speech and Nonspeech Signals*

David B. Pisoni
Indiana University
Bloomington, Indiana 47405

*This is the final draft of a paper presented as the introduction to Symposium No. 8 "Perception of Speech versus Non-Speech" held at the Ninth International Congress of Phonetic Sciences, Copenhagen, August 6 to 11, 1979.

Some Remarks on the Perception of Speech and Nonspeech Signals

Historically, the study of speech perception may be said to differ in a number of ways from the study of other aspects of auditory perception. First, the signals used to study the functioning of the auditory system were simple and discrete, typically varying along only a single physical dimension. In contrast, speech signals display very complex spectral and temporal relations. Although speech signals have also been varied along single physical dimensions, the perceptual consequences of such manipulation have not always followed from "equivalent" stimulations of a nonspeech nature. Alternatively, we may presume that the complexity of the spectral and temporal structure of speech and its variaton is one additional source of perceptual differences between speech and nonspeech signals.

Second, most of the research dealing with auditory psychophysics that has accumulated over the last thirty years has been concerned with the discriminative capacities of the sensory transducer and the functioning of the peripheral auditory mechanism. In the case of speech perception, however, the relevant mechanisms are assumed to be centrally located and intimately related to more general cognitive processes that involve the encoding, storage and retrieval of information in memory. Moreover, experiments in auditory psychophysics have typically focused on experimental tasks and paradigms that involve discrimination rather than identification or recognition, processes thought to be most relevant to speech perception. All in all, it is generally believed that a good deal of what has been learned from research in auditory psychophysics and general auditory perception is only marginally relevant to the study of speech perception and to an understanding of the underlying perceptual mechanisms. This situation has changed for the better in recent years as shown by the work of Dr. Divenyi and other psychophysicists who have become concerned with questions of speech perception.

Despite these obvious differences, investigators have been quite interested in the differences in perception between speech and nonspeech signals. That such differences might exist was first suggested by the report of the earliest findings of categorical discrimination of speech by Liberman et al. (1957). And it was with this general goal in mind that the first so-called "nonspeech control" experiment was carried out by Liberman et al. (1961) in order to determine the basis for the apparent distinctiveness of speech sounds. In this study the spectrographic patterns for the /do/ and /to/ continuum were inverted producing a set of nonspeech patterns that differed in the onset time of the individual components. The results of perceptual tests showed peaks in discrimination for the speech stimuli replicating earlier findings on categorical perception. However, there was no evidence of comparable discrimination peaks for the nonspeech stimuli, a result that was interpreted at the time as further evidence for the distinctiveness of speech sounds and the effects of learning on speech perception.

Numerous speech-nonspeech comparisons have been carried out over the years since these early studies, including several of the contributions to the present symposium. For the most part, these experiments have revealed results quite similar to the original findings of Liberman et al. Until quite recently, research reports have confirmed that performance with nonspeech control signals failed to show the same discrimination functions that were observed with the parallel set of speech signals (Cutting and Rosner, 1974; Miller et al., 1976; Pisoni, 1977). Subjects typically responded to the nonspeech signals at levels approximating chance performance. In more recent years, such differences in perception have been assumed to reflect two basically different modes of perception--a "speech mode" and an "auditory mode." Despite some attempts to dismiss this dichotomy, additional evidence continues to accumulate as has been suggested by several of the new findings summarized in the papers included in this symposium.

The picture is far from clear, however, for the problems inherent in comparing speech and nonspeech signals have generated several questions about the interpretation of results obtained in earlier studies. First, there is the question of whether the same psychophysical properties found in the speech stimuli were indeed preserved in the parallel set of nonspeech control signals. Such a criticism is appropriate for the original /do/--/to/ nonspeech control stimuli which were simply inverted patterns reproduced on the pattern playback. The same remarks also apply to the well-known "chirp" and "bleat" control stimuli of Mattingly et al. (1971) which were created by removing the formant transitions and steady-states from the original speech context. These stimuli were presented in isolation to subjects for discrimination. Such manipulations, while nominally preserving the phonetic "cue" obviously result in marked changes in the spectral context of the signal which no doubt affects the detection and discrimination of the original formant transition. Such criticisms have been taken into account in the more recent experiments comparing speech and nonspeech signals as summarized by Dr. Dorman and Dr. Liberman, in which the stimulus materials remain identical across different experimental manipulations.

While these more recent studies relieve some of the ambiguities of the earlier experiments, problems still remain in drawing comparisons between speech and nonspeech signals. For example, subjects in these experiments rarely practice with the nonspeech control signals to develop the competence required to categorize them consistently. With complex multidimensional signals it is quite difficult for subjects to attend to the relevant attributes that distinguish one signal from others presented in the experiment. A subject`s performance with these nonspeech signals may therefore be no better than chance if he/she is not attending selectively to the same specific criterial attributes that distinguished the original speech stimuli. Indeed, not knowing what to listen for may force a subject to selectively attend to an irrelevant or misleading attribute of the signal itself. Alternatively, a subject may simply focus on the most salient auditory quality of the perceived stimulus without regard for the less salient acoustic properties which often are the most important in speech such as burst spectra or formant transitions. Since almost all of the nonspeech experiments conducted in the past were carried out without the use of discrimination training and feedback to subjects, an observer may simply focus on one aspect of the stimulus on one trial and an entirely different aspect of the stimulus on the next trial. Without training experience to help the subject identify the criterial properties, the observed performance may be close to chance, a result that has been reported quite consistently in the literature.

Setting aside some of these criticisms, the question still remains whether drawing comparisons in perception between speech and nonspeech signals will yield meaningful insights into the perceptual mechanisms deployed in processing speech. In recent years, the use of cross-language, developmental and comparative designs in speech perception research has proven to be quite useful in this regard as a way of separating out the various roles that genetic predispositions and experience play in speech perception. On the one hand, these types of investigations provide needed information about the course of learning and perceptual development since spoken language must be acquired in the local environment through social contact. On the other hand, comparative studies with both speech and nonspeech stimuli are useful in defining the lower limits on auditory system function. However, there are serious limitations in studies of this kind. For example, while it is cited with increasing frequency that chinchillas categorize synthetic stimuli differing in VOT in a manner quite similar to <u>English-speaking</u> adults, little if anything is ever mentioned about the chinchilla`s <u>failure</u> to carry out the same task with stimuli differing in the cues to place of articulation in stops, a discrimination that even young prelinguistic infants can make (Eimas, 1975). Should we then conclude that the English voicing contrast is purely sensory in origin, while place of articulation or voicing in Thai is somehow more "linguistic," brought on by inheritance or very early experience? With a little reflection, I think the answer must surely be negative.

Such comparative studies are useful in speech perception research only to the extent that they can specify the lower-limits on the sensory properties of the stimuli themselves. However, these findings are incapable, in principle, of providing any further information about how these signals might be "interpreted" or coded within the context of the experience and history of the organism. Animals simply do not have spoken language and they do not and cannot recognize, as far as I know, differences between phonetic and phonological structure, a fundamental dichotomy in all natural languages.

Cross-language and developmental designs have also been quite useful in providing new information about the role of early experience in perceptual development and the manner in which selective modification or tuning of the perceptual system takes place. Although the linguistic experience and background of a listener was once thought to tightly control his/her discriminative capacities in speech perception experiments, recent findings strongly suggest that the perceptual system has a good deal of plasticity for retuning and realignment, even into adulthood. The extent to which control over the productive abilities remains plastic is still a topic to be explored in future research.

To what extent is it then useful to argue for the existence of different modes of perception for speech and nonspeech signals? Some investigators such as Dr. Ades would simply dismiss the distinctions drawn from earlier work on the vague grounds of parsimony and generality. He has argued recently (Ades, 1977) and in his contribution to this symposium that differences in perception between speech and nonspeech or consonants and vowels can be accounted for simply by recourse to the notion of "range" or the width of the context expressed in terms of the number of JNDs. As long as the range is small, absolute identification performance will be as good as differential discrimination. When the range is large, however, discrimination will be better than identification. Thus according to the account offered by Ades, a consonant continuum should display a smaller range than a vowel continuum. But as shown in Slide 1 the facts are quite the reverse of his predictions.

May I Have Slide 1 Please?

In this figure we have reproduced the identification data collected by Perey and Pisoni (1977) in a magnitude estimation task. On each trial subjects had to respond to a stimulus with a rating on a scale from 1 to 7. One group of subjects received a consonant continuum differing in VOT, another received a vowel continuum. Through various transformations of the obtained stimulus-response matrix, scale scores were derived and an estimate of the perceived psychological spacing between stimuli was obtained. Scale scores are expressed in this figure in terms of d`s and by summing these individual values, an estimate of the total range or spacing of the stimuli was obtained. The cumulative d` is shown on the far right of each panel. Notice that the cumulative d` for the vowels shown on the top is 4.19 while the value for the consonants shown on the bottom is 4.28.

If stimulus range were the correct explanation of the differences in perception between consonants and vowels as Dr. Ades would have us believe, the consonants should have displayed the smaller range. Obviously, this is simply not the case. However, what is of interest in this figure is the psychological spacing of signals within each panel. For the consonants, the spacing between adjacent stimuli is clearly unequal with a grouping close to the endpoints of the series. For the vowels, the spacing is more nearly equal across all the test stimuli suggesting the possibility of better resolution in discrimination, a result that has been known for many years. Thus, Dr. Ades` argument that the range of stimuli can account for differences in perception between consonants and vowels or speech and nonspeech would seem to be incorrect, despite his attempts to generalize the Durlach and Braida (1969) model to speech perception. Moreover, this is a curious position to maintain anyway as it is commonly recognized, not only in speech perception research but in other areas of perceptual psychology, that "nominal" stimuli may receive differential amounts of processing or attention by the subject, that subjects may organize the interpretation of the sensory information differently under different conditions and that the sensory trace of the initial input signal may show only a faint resemblance to its final internal representation resulting from encoding and storage in memory.

It is hard to deny that a speech signal elicits a characteristic mode of response in a human subject--a response that is not simply the consequence of an acoustic waveform leaving a meaningless sensory trace in the auditory periphery. Nevertheless, there is a great deal to learn about how the auditory system codes complex acoustic signals such as speech. Dr. Dorman, in summarizing work on the perception of transitions in speech and nonspeech context, has tried to establish the need for a specialized speech processor to account for differences in labelling of sine-wave stimuli when heard as either speech or nonspeech. Such explanations seem to me entirely premature at this time as the relevant psychophysical experiments with nonspeech signals have simply not been carried out yet. To remedy this state of affairs we have begun to collect labelling data in our laboratory recently using brief FM stimuli followed by a constant frequency (CF) steady-state. Schematized spectrograms of the test stimuli are shown in Slide 2.

May I Have Slide 2 Please?

The left panel of this figure shows an idealized set of stimuli differing in the initial starting frequency of the FM. Three steady-state (CF) frequencies were selected, 850, 1500 and 2300 Hz. For each set we generated 21 test signals which spanned a range of 500 Hz above and below the CF of the steady-state component. In Experiment I the three sets of signals consisted of an isolated single component as shown on the left. In

Experiment II we added an additional 500 Hz component to each of the original three sets of stimuli.

Subjects were required to identify the stimuli as "rising," "level" or "falling" after a brief training period with good exemplars selected from each category. The results of both experiments are shown in Slide 3.

May I Have Slide 3 Please?

The labelling functions shown at the top for the three CF conditions reveal that the middle or "level" category response increases slightly in size as the CF of the steady-state increases from 850 Hz to 1500 Hz, a result that is consistent with what is known about frequency resolution in the auditory system. Over a wide range of frequencies, discrimination follows Weber`s law. Thus, the level category should widen as the frequency of the steady-state increases for the same difference in initial starting frequency. Note that we have plotted starting frequency on a linear rather than log scale.

The results for Experiment II in which an additional steady-state component was added are shown in the lower panel of the figure. Notice that for the 850 Hz condition the "level" category is now slightly larger than in the top panel suggesting the strong possibility of some interaction between the individual components. However, the other two conditions in Experiment II show a somewhat narrower range for the "level" category compared to the top panel indicating better resolution of frequency in the presence of another signal, a well known fact in auditory psychophysics.

These recent findings were not originally intended to refute the arguments of Dorman and his colleagues who favor the postulation of some specialized perceptual mechanism for processing speech signals. Rather, I simply wanted to illustrate by way of example that the location of perceptual categories observed with nonspeech signals is not rigidly controlled by some simple physically defined invariant such as the direction of the frequency change. Moreover, as Dr. Divenyi has pointed out so well in his paper, we need to know much more about how the basic constraints of the auditory system affect the way speech is initially coded for subsequent processing. Thus, in the present case several basic facts about frequency discrimination are sufficient to account for changes in our subjects perceptual categorization of nonspeech FM`s that are similar to speech. Whether it will be possible to generalize such psychophysical explanations to more complex signals such as speech remains to be seen from future research currently in progress in our laboratory and elsewhere.

In summary, there still appears to be good evidence for distinguishing between speech and nonspeech signals and for recognizing the existence of two distinct modes of perception, one associated with the sensory or psychophysical correlates of acoustic signals and the other with the interpretation and coding of acoustic signals as speech. Recent work has attempted to make these differences more precise by subjecting them to experimental test and searching for common underlying explanations.

Taken together such results suggest to me that, just as in the case of "species-typical responding" observed in the behavior of other animals, the notion of a "speech mode" of perception captures certain aspects of the way human observers typically respond to speech signals that are highly familiar to them. We still do not know if it is simply a matter of familiarity such as with music or whether there is something deeper and more closely related to biological survival of the organism. Nevertheless, such a conceptualization does not, at least in my view, commit one to the view that human listeners cannot respond to speech signals in other ways more closely correlated with the sensory or psychophysical attributes of the signals themselves. To deny the speech mode, however, is to ignore the fact that acoustic signals generated by the human vocal tract are used in a distinctive and quite systematic way by both talkers and listeners to communicate linguistically, a species-typical behavior that is restricted, as far as I know, to *Homo sapiens*.

Past experiments comparing the perception of speech and nonspeech signals have been quite useful in characterizing how the phonological systems of natural languages have, in some sense, made use of the general properties of sensory systems in selecting an inventory of phonetic features and their acoustic correlates (Stevens, 1972). The relatively small number of distinctive features and their acoustic correlates that can be observed across a wide variety of diverse languages implies that there is a common sensory basis for language production, a common means of controlling the mechanisms of speech production and a common cognitive definition of linguistic structure. Whether these facts are causally related will no doubt be a matter of much debate, speculation and new research in the years to come. It is clear, nevertheless, that the distinctions drawn in perception between speech and nonspeech signals still remain fundamental ones setting apart research on speech perception from the study of auditory psychophysics and the field of auditory perception more generally.

Acknowledgements

The preparation of this paper was supported, in part, by NIMH grant MH-24027 and NINCDS grant NS-12179 to Indiana University in Bloomington. I am grateful to Peter Jusczyk and Jim Sawusch for comments on an earlier draft of the paper. Robert Remez discussed many of the theoretical issues summarized in the paper with me at length and provided helpful editorial comments that improved the overall exposition and quality. His help is greatly appreciated.

References

Ades, A. E. (1977) "Vowels, consonants, speech and nonspeech," Psych. Rev. 84, 524-530.

Cutting, J. E. and Rosner, B. S. (1974) "Categories and boundaries in speech and music," Perc. Psych. 16, 564-570.

Durlach, N. I. and Braida, L. D. (1969) "Intensity perception I. Preliminary theory of intensity resolution," JASA 46, 372-383.

Eimas, P. D. (1974) "Auditory and linguistic processing of cues for place of articulation by infants," Perc. Psych. 16, 513-521.

Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. (1957) "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psych. 54, 358-368.

Liberman, A. M., Harris, K. S., Kinney, J. A. and Lane, H. L. (1961) "The discrimination of relative onset time of the components of certain speech and non-speech patterns," J. Exp. Psych. 61, 379-388.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K. and Halwes, T. G. (1971) "Discrimination in speech and non-speech modes," Cogn. Psych. 2, 131-157.

Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J. and Dooling, R. J. (1976) "Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception," JASA 60, 410-417.

Perey, A. J. and Pisoni, D. B. (1977) "Dual processing versus response-limitation accounts of categorical perception: A reply to MacMillan, Kaplan and Creelman," JASA 62, S1, 60-61.

Pisoni, D. B. (1977) "Identification and discrimination of the relative onset of two component tones: Implications for voicing perception in stops," JASA 61, 1352-1361.

Stevens, K. N. "The quantal theory of speech: Evidence from a articulatory-acoustic data." In E. E. David, Jr. and P. B. Denes (Eds.) Human Communication: A Unified View. New York: McGraw-Hill, 1972.

Figure Captions

Figure 1. Scale values showing the perceived psychological space for consonants and vowels. Data were taken from Perey and Pisoni (1977) who required subjects to use a rating response in identification.

Figure 2. Schematized patterns showing the time course of the nonspeech FM stimuli: The panel on the left illustrates the test stimuli without spectral context, the panel on the right shows the addition of a low frequency component to the same signals.

Figure 3. Identification data for FM stimuli obtained with three different steady-state CF`s, 850 Hz, 1500 Hz and 2300 Hz. The top panel shows the identification data collected for FM`s without context, the lower panel shows the data for test signals with the additional steady-state context present.

SCALE VALUES & RESPONSE CRITERIA FROM IDENTIFICATION

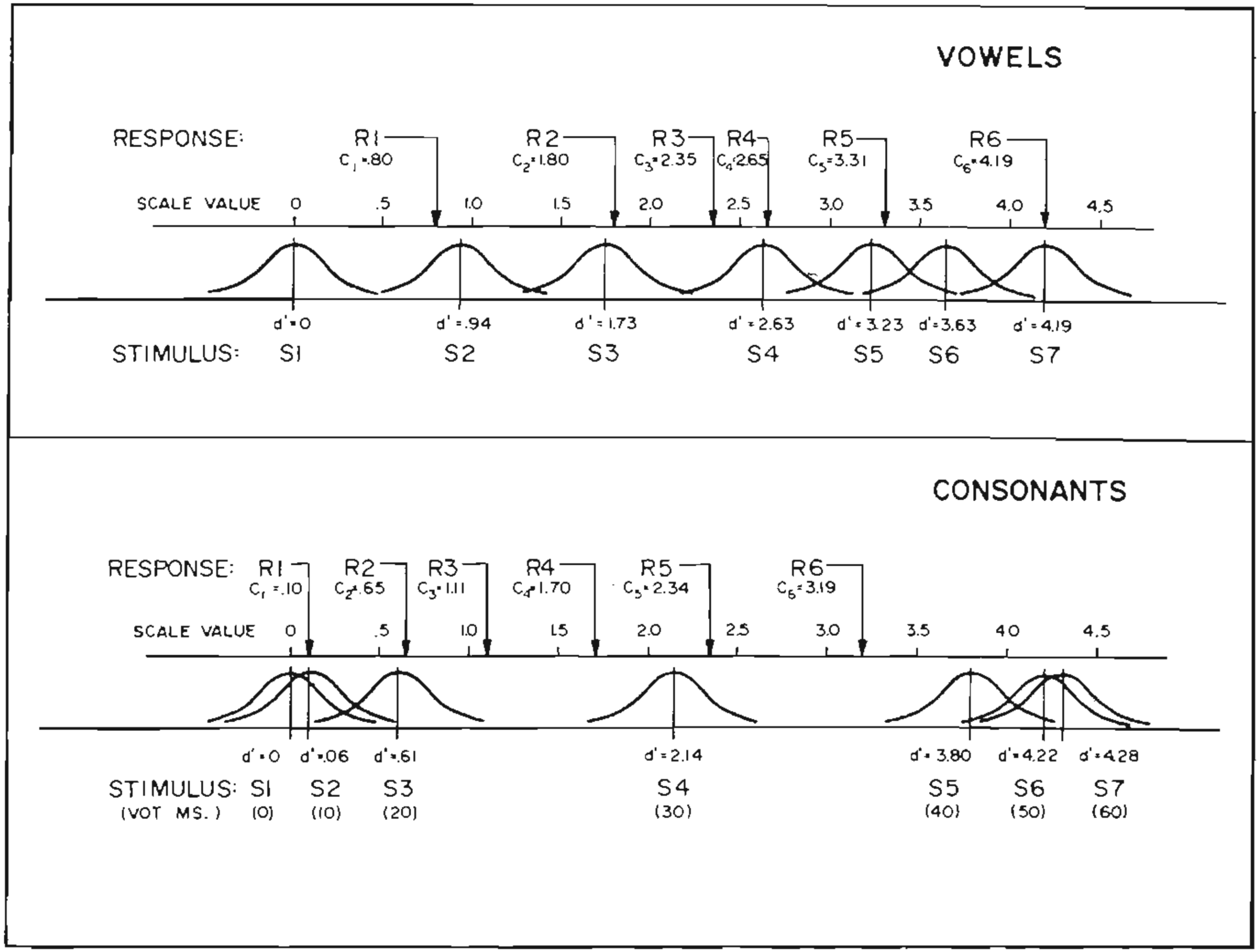


Figure 1.

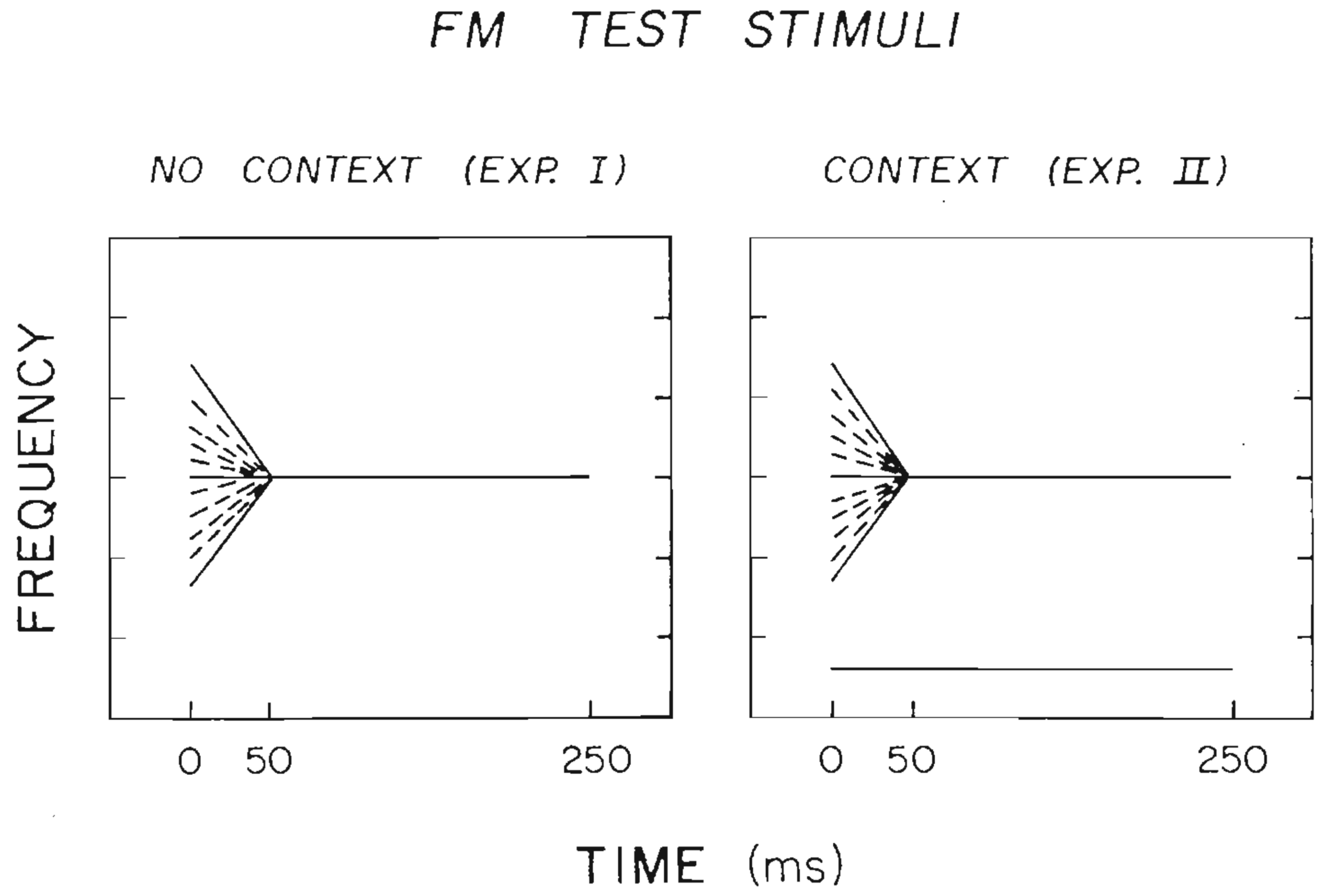

FM TEST STIMULI
NO CONTEXT (EXP. I)
CONTEXT (EXP. II)
FREQUENCY
0
50
250
0
50
250
TIME (ms)

Figure 2.

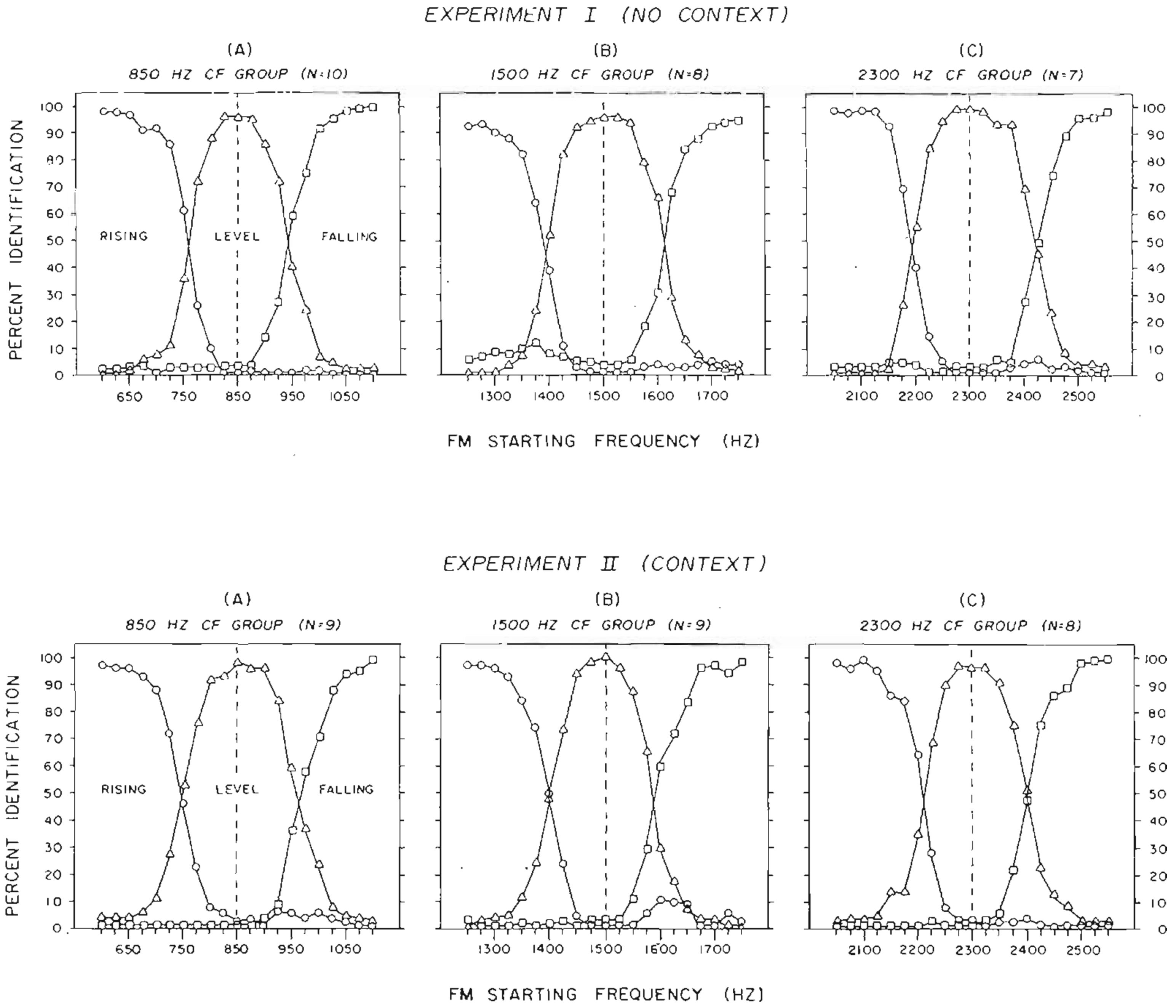

EXPERIMENT I (NO CONTEXT)
(A)
850 HZ CF GROUP (N=10)
(B)
1500 HZ CF GROUP (N=8)
(C)
2300 HZ CF GROUP (N=7)
PERCENT IDENTIFICATION
RISING
LEVEL
FALLING
100
90
80
70
60
50
40
30
20
10
0
650
750
850
950
1050
1300
1400
1500
1600
1700
2100
2200
2300
2400
2500
FM STARTING FREQUENCY (HZ)
EXPERIMENT II (CONTEXT)
(A)
850 HZ CF GROUP (N=9)
(B)
1500 HZ CF GROUP (N=9)
(C)
2300 HZ CF GROUP (N=8)
PERCENT IDENTIFICATION
RISING
LEVEL
FALLING
FM STARTING FREQUENCY (HZ)

Figure 3.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 5 (1978-79) Indiana University]

Continuous Spectral Change as Acoustic Cues to Place of Articulation

Diane Kewley-Port

Two experiments were carried out to examine the role of change in spectral prominences over time as cues to place in several vowel contexts. Spectral sections were calculated every 5ms using linear prediction analysis for five repetitions of the stops /b,d,g/ before /i,I,e^I,ε,ae,a/ for one talker. In the first experiment voiced formant transitions were measured. Results indicated that formant transitions were not sufficient cues to place for all vowel contexts. A second experiment established a well-defined spectral continuity between burst and the early portions of the formant transitions using three-dimensional running spectral displays. Three features having acoustic descriptions related to the theoretical output of the vocal tract were defined. Using these features subjects identified 88% of the consonants correctly from the visual displays.

Introduction

The search for the cues to place of articulation in stop consonants has usually focused on two acoustically distinct segments, the release burst and the voiced formant transitions into the vowel. While some research has suggested that the burst alone might be an invariant place cue , most investigators have concluded that a combination of burst and formant transitions is necessary for recognition. In general, the burst has been characterized by a single spectral section of the voiceless waveform, while formant transitions have been defined as the change in spectral prominences over time after voicing onsets. Two recent studies have employed somewhat different analyses. Stevens and Blumstein (1978) have provided some evidence for invariant burst onset spectra which average both burst and early portions of the transitions in a single 26ms window. Searle et al. (1979), on the other hand, have examined displays of the spectral prominences of stops continuously from the burst into the vowel in a series of closely spaced running spectra. However, the features they defined as cues to place were not spectrally changing parameters, but were measurements of four spectral sections at least 20ms apart in the CV. Their features correctly classified stops for place about 79%.

The present investigation focused on change in spectral energy over time from the point of view that spectral change may be a better representation of the information output from the peripheral auditory system. Experiment 1 reexamines the role of formant transitions as cues to place. Although formant transitions are often assumed to be sufficient cues to place based on speech synthesis (Liberman et al., 1954) and tape splicing experiments (Dorman et al., 1977; Just et al., 1978) very few detailed measurements of formant transitions have been obtained from real talkers (Fant, 1973; Ohmann, 1968). Experiment 1 examines the potential role of formant transitions as cues to place from measurements of one talker. The outcome of Experiment 1 suggested that the measured transitions do not separate the classes of stop consonants very well. In Experiment 2, running spectra which included both the burst and transitions were examined for place cues.

Experiment 1

Method. The test syllables consisted of the voiced stops /b,d,g/ followed by /i,I,e^{I},ε,ae,a/. These 18 CV's were embedded in the carrier sentence "Teddy said CV", and were presented under computer control on a CRT monitor. One male speaker of American English , a phonetician, recorded 10 randomizations of the sentences. Five of these lists were digitized by a 12 bit A/D at a 10K sampling rate and stored on disk. The linear prediction analysis algorithms of Markel and Gray (1976) were used to calculate the formants. Fourteen linear prediction coefficients were calculated for the first 100ms of each CV at 5ms intervals called frames. A 20ms Hamming window was applied and the window of the first frame was centered over the burst. Formant transitions were located manually using strict definitions of voicing onset and formant steady-state with the assistance of computer driven visual displays. The voicing onset was determined by visually examining the smoothed spectral sections of each frame following the burst. The presence of a sharp F1 spectral peak accompanied by an abrupt increase in RMS energy was used to determine the first voiced frame. Vowel steady-states were determined from the CRT displays of the formants and a listing of the computed formant values. Vowel steady-state was defined for each formant as the frame in which either the frequency change fell to less than 10 Hz per 5ms frame, or where the formant back-tracked to the same value within 4 frames. Figure 1 shows typical markings for the steady-states.

Insert Figure 1 about here

Measurements of onset frequency and frame, steady-state frequency and frame, as well as any deviations from approximations to straight line segments were always recorded for the first 3 formants, and for the fourth formant when present. In four CV's out of 90, the first three formants were not consistently tracked by the algorithm, even when the number of coefficients was changed. In those cases, another repetition of that CV from the original recording was substituted.

Results and Discussion. Comparisons of the individual plots of the five repetitions for each CV revealed that each transition could be approximated by a straight line. Measurements for F1, F2 and F3 were averaged over the five repetitions and analyzed statistically with a one-way analysis of variance and the Scheffe posthoc analysis to test for significant differences between all pairs of stops.

The first question asked was whether the vowel steady-states showed context effects due to the consonants. Table 1 summarizes the results of the statistical analyses.

Insert Table 1 about here

No significant differences were observed across consonants for F1; only 4 differences out of 18 were found for F2, and only 1

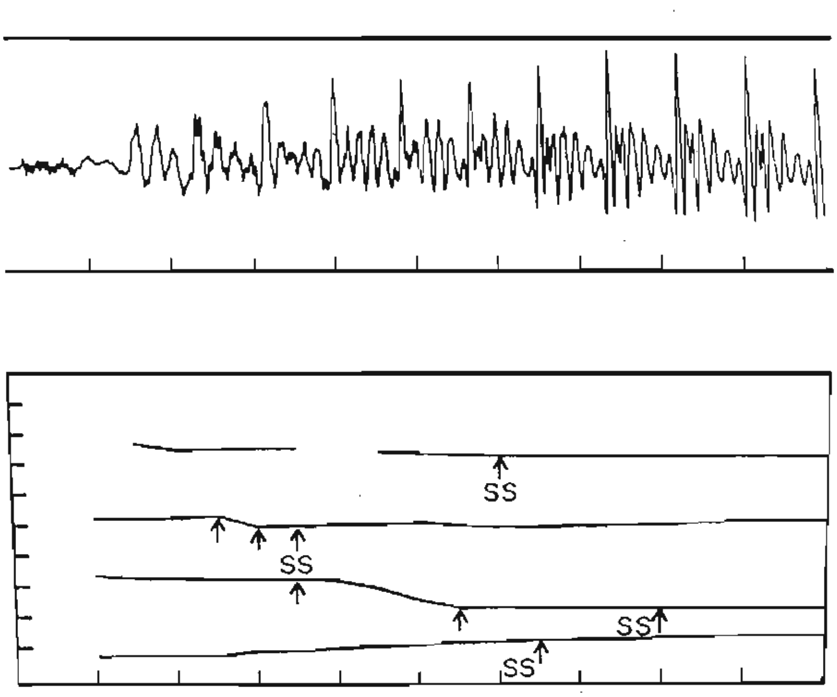


Figure 1. Formant transitions for /da/ aligned with the waveform. "SS↑" indicates the frame in which the formant steady state was acheived. "↑" alone denotes frames where transitions deviated from straight line approximations.

	i	I	e^{I}	ε	ae	a
F1	(bd)(dg)	(bdg)	(bd)(dg)	(bdg)	(bdg)	(bdg)
F2	(bdg)	(bd)g	(bd)g	(bd)(dg)	b(dg)	b(dg)
F3	(bd)g	(bdg)	(bdg)	(bdg)	(bdg)	(bdg)

TABLE 1. Statistical groupings of steady-state frequencies for each vowel formant. Stops inside parentheses indicate the consonantal contexts for each steady state formant which were grouped together as not statistically different.

difference was observed for F3. Thus, the vowel targets produced by this speaker were quite similar across the initial stops, suggesting that differences in vowel target frequencies alone could not serve as places cue for /b,d,g/.

The next question sought to determine if the formant frequencies at onset when measured from real speech could be sufficient cues to place. Table 2 summarizes the results of the statistical analyses.

Insert Table 2 about here

We assume that onset frequency of the formant transitions could be a cue to place only if the measured values are significantly different from one another. The results showed that Fl is not a possible cue to place for any vowel context. Only one vowel, /ae/, showed complete separation of /b,d,g/ onsets for the frequencies of both F2 and F3. On the other hand, only one vowel, /i/, had overlap of onset frequencies - for (b,d) - for both F2 and F3. For vowels /i,e^{I},ε /, F2 onsets were significantly different, but not F3. Finally, /a/ presented a confusing picture with /b/ onsets distinct from /d,g/ for F2, whereas /d/ onsets were distinct from /b,g/ for F3. From these results, we conclude that frequency onsets for Fl, F2 and F3 alone are not sufficient cues to place for the stops in these vowels.

Onset frequencies were analyzed further to determine if a two dimensional, F2 X F3 space could be used to classify the stops.

	i	I	e^I	ε	ae	a
F1	(bdg)	b (dg)	b (dg)	(bdg)	b (dg)	(bdg)
F2	(bd) g	*	*	*	*	b (dg)
F3	(bd) g	b (dg)	(bdg)	b (dg)	*	(bg) d

TABLE 2. Statistical groupings of onset formant frequencies. "*" indicates that the onset frequencies for /b,d,g/ were all significantly different from one another. Stops inside the parentheses indicate which onsets were grouped together as not significantly different.

Figure 2 shows the formant values plotted in a linear F2 X F3 space.

Insert Figure 2 about here

Discriminant Analysis was used to calculate an optimal F2 X F3 space separately for each vowel. Stops were correctly assigned for place 100% except for one /bi/ classified as /di/, and one /bI/ classified as /dI/. Thus, F2 and F3 onset frequencies defined in a two-dimensional, vowel context-dependent space can be very effective cues to place. The extent to which human speech perception mechanisms can make use of vowel-dependant, two-dimensional representations of formant frequencies to identify place of articulation remains to be seen in future work.

Durations of the formant transitions were also calculated, averaged and analyzed as place cues separately for each vowel. Durations varied greatly from token to token and formant to formant. The standard deviations were often as great as the transition durations themselves. The statistical analysis showed that only the F1 transition duration for the vowel /ae/ successfully sorted the stops for place. Thus, for these vowels there were no consistent differences in the transition durations which might serve as place cues.

Finally, voice onset time was examined as a possible cue to place of articulation. VOT was measured here as the difference between the frame in which voicing onsets, according to the

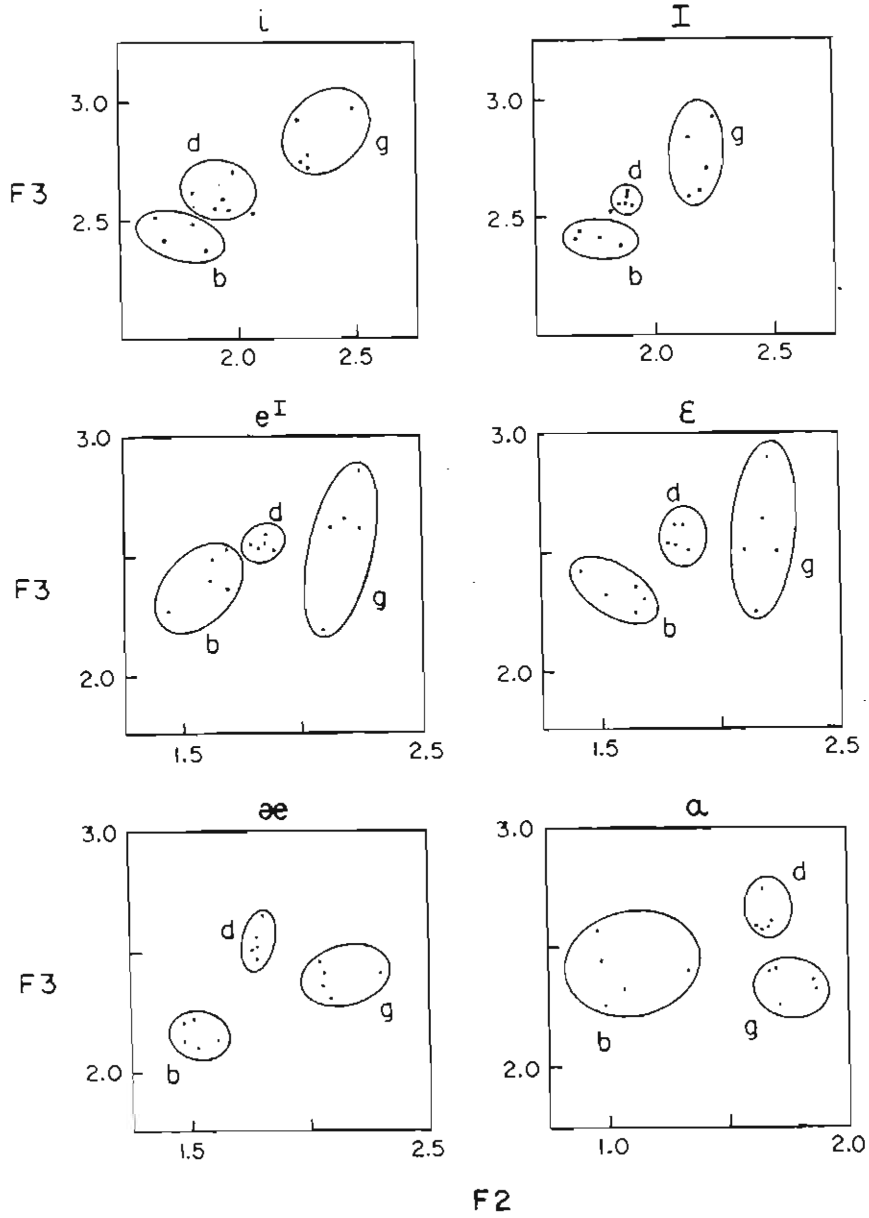


Figure 2. Each panel shows the groupings of onset formant frequencies produced by Discriminant Analysis using standard F2 X F3 coordinates. For the vowels /i/ and /I/, one point each measured froma /b/ was classified as a /d/. The axes for F2 and F3 are linear, and the ticks are for KHz.

criteria mentioned above, and the first frame which contained the release burst. Since it was obvious that VOT could reliably classify stops for each vowel, VOT statistics were collasped across all six vowel contexts.

Insert Table 3 about here

Table 3 shows the results where VOT has been converted to milliseconds. The average VOT values measured for this speaker showed the characteristic differences in VOT for different places of articulation, namely that VOT is longer as place of articulation goes from the front of the mouth (labial) to the back (velar) (Lisker and Abramson, 1967). The statistical analysis showed that VOT alone was highly effective in classifying place of articulation. A Discriminant Analysis was then used to produce a confusion matrix for stop classification. Table 3 shows that most errors occur in classifying /d/. The overall percent correct classification is 89%. Although these results are for one subject speaking at a constant tempo, VOT is a very effective cue for classifying stop consonants, and is, moreover, context invariant. More will be said about VOT in Experiment 2.

Conclusions. Detailed measurements of the formant transitions from one talker showed that the only distinctive properties which could serve as cues to place of articulation were the onset frequencies for F2 and F3. However, onset frequencies were only distinctive for the vowel /ae/, and could not reliably

	Measured or Predicted		
	b	d	g
VOT	3.1ms	11ms	18.5ms
b	n=30 100%	n=0 0%	n=0 0%
d	n=1 3.3%	n=22 73.3%	n=7 23.3%
g	n=0 0%	n=2 6.7%	n=28 93.3%

TABLE 3. VOT is the average VOT for each stop calculated across all 6 vowels. Below is the confusion matrix produced by Discriminant Analysis for assigning stop categories across all vowels based on VOT alone.

differentiate place for /i/ or /a/. Representations of onset frequencies in a two-dimensional F2 X F3 space were effective in classifying place of articulation. Utilization of these two-dimensional representations in human perception remains to be established. On the basis of the absence of any differences between stops in transition durations or vowel steady-states, and the unreliability of F2 and F3 onsets to be distinctive cues even when vowel context is known, we conclude that formant transitions alone are not likely to be sufficient cues to place of articulation.

Experiment 2

Introduction. Since reliable cues to place of articulation derived from spectrally changing parameters were difficult to find in the transitions alone, it seemed reasonable to look at burst and transitions together. This seemed especially important since Stevens and Blumstein (1978) claimed that invariant cues for place across any vowel context could by located in the first 26ms of spectral energy integrated over burst and formant transitions. However, the recent paper by Searle et al. (1979) provided the inspiration to create running spectral displays which demonstrate changes in spectral energy over time in contrast to the single integrated spectrum used by Stevens and Blumstein (1978).

The spectral sections from the linear prediction analysis calculated in Experiment 1 were plotted in a three-dimensional running spectra displays as shown in Figure 3.

Insert Figure 3 about here

The first frame of each display is the burst spectrum where the burst onset was centered in the Hamming window such that the effective windowing was 5ms. Each display includes 8 frames, or 40ms, of the CV syllable. As can be seen, these displays show continuity of change in the spectral prominances from the release burst through the early portion of the voiced formant transitions.

Visual examination of the 90 CV's in Experiment 1 showed that there might be some features in the running spectra that classified displays for place in a context-invariant way. A pilot study suggested the following three binary features in conjunction with a feature matrix could be used to classify the stops. Figure 1 provides examples of these features.

Feature 1) *Tilt of the spectrum at burst onset.* Tilt is estimated by visually fitting a straight line through the first frame of the running spectrum. The feature categories were R = rising and F = flat, falling or ambiguous. Rising burst spectra are characteristic of alveolar articulation, falling of labial articulation, and velar bursts have not been generally characterized as rising or falling (Fant, 1960; Stevens and Blumstein, 1978).

Feature 2) *Late onset of low frequency energy*. Late onset is defined as the occurence of high amplitude, low frequency peaks starting in the fourth frame or later. Feature categories were L =

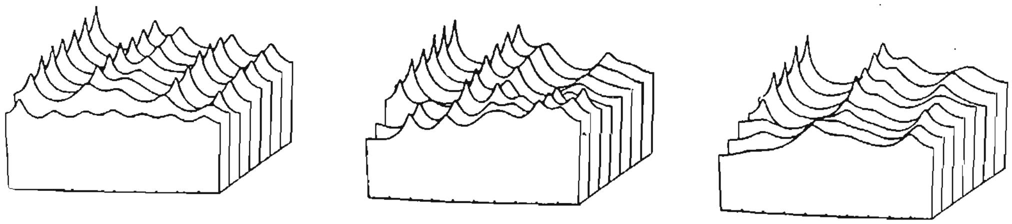


Figure 3. Running spectral displays beginning with the burst. The x-axis is linear frequency from 1 to 5000 Hz; the y-axis is relative dB; and the z-axis is time, 5ms per spectral section. Fig.3a displays [be^I], Fig.3b is [de^I], and Fig.3c is [ge^I].

late onset and N = no late onset. As mentioned in Experiment 1, this is a measure of VOT which was very effective in classifying place for this subject. The concept of Late onset is directly related to the fact that /g/ has a longer VOT than /b/ or /d/, and is usually greater than 20ms.

Feature 3) Mid-frequency peaks extending over time. This feature involves a single, prominent peak falling between 1000 and 2500 Hz that occurs on three or more frames. The feature categories were Y = yes, peaks exists and N = no such peaks are present. These peaks reflect the compact spectrum characteristic of velar consonants which is associated with the resonant cavity in front of the constriction (Fant, 1960; Stevens and Blumstein, 1978).

The feature matrix constructed to assign consonants is shown in Table 4. A study was then conducted to validate the effectiveness of these features for classifying stops.

Insert Table 4 about here

Method. Slides were made of the running spectra for all 90 CV's. Five slides, including Figures 3a, 3b, and 3c, were duplicated as examples. Six subjects who had had a phonetics course, but no familiarity with this form of analysis or running spectral displays, participated in the experiment. Written instructions included the feature definitions and feature matrix similiar to that described above. The five examples were

Tilt	Late Onset	Mid-freq. Peaks	Assigned Consonant
F	N	N	b
R	?	N	d
?	L	Y	g

TABLE 4. Feature matrix to assign consonants. "?" means that either feature category may occur for that stop.

discussed. Subjects then viewed a random ordering of the slides and wrote down the feature category responses, and the consonant for each display.

Results. The feature assignments showed 95% agreement between subjects. Correct identification of the consonants was 88%. Errors were not evenly distributed for place; /b/ had 1% errors, /d/ had 10% errors and /g/ had 21% errors. Subsequent analysis showed that the /g/ errors occurred primarily on four slides where the prominant mid-frequency peak was not a " single" peak. Overall correct identification improves to 93% if these four slides are excluded. Thus a slightly different wording in the feature definitions would presumably increase the correct identification of /g/.

Conclusion. Features associated with displays showing the continuity of spectral change from the burst into the early vocal portions of a CV look promising as a way of representing cues to place of articulation in stop consonants. The features defined here are not arbitrary, but relate directly to the theoretical output of the vocal tract. Experiment 2 is in the nature of a pilot study and more talkers using more vowels should be studied before the effectiveness of this approach can be validated. However, both experiments together demonstrate clearly that in natural speech formant transitions and bursts should not be treated as separate cues to place of articulation. Although they are acoustically distinct segments, they have spectrally continuous properties which presumably result from a unifying articulatory gesture.

Acknowledgements

This research was supported by NINCDS Grant NS-12179 to Indiana University. I am indebted to Professor David B. Pisoni for his advice and consultation in many phases of this research, as well as his comments in preparing this manuscript.

References

Dorman, M., Studdert-Kennedy, M. and Raphael, L. (1977), "The invariance problem in initial voice stop consonants: Release bursts and formant transitions as functionally equivalent context-dependent cues", Percept. Psychophys. 22,109-122.

Fant, G. (1960), Acoustic Theory of Speech Production., 's-Gravenhage:Mouton.

Fant, G. (1973) Stops in CV-syllables. In G. Fant, Speech Sounds and Features, Cambridge, Mass.: MIT, 110-142.

Just, M., Suslick, R.L., Michaels, S. and Shockey, L. (1978), "Acoustic cues and psychological processes in the percepion of natural stop consonants, Percept. Psychophys. 24, 327-336.

Liberman, A.M., Delattre, P.C., Cooper, F.S. and Gerstman, L.H. (1954), "The role of consonant-vowel transitions in the perception of the stop and nasal consonants", Psych. Mono. 68, 1-13.

Lisker, L. and A.S. Abramson (1967), "Some effects of context on voice onset time in English stops", Lang. & Speech 10, 1-28.

Markel, J.D. and Gray, A.H. (1976), Linear Prediction of Speech, New York:Springer-Verlag.

Ohmann, S.E.G. (1968) Coarticulation in VCV utterances: Spectrographic measurements. J. Acous. Soc. Am. 39, 151-168.

Searle, C.L., Jacobson, J.Z. and Rayment, S.G. (1979), "Stop consonant discrimination based on human audition", J. Acous. Soc. Am. 65, 799-809.

Stevens,K.N. and Blumstein, S.E. (1978), "Invariant cues for place of articulation in stop consonanats", J. Acous. Soc. Am. 64 ,1358-1368.

[RESEARCH ON SPEECH PERCEPTION Progress Report No. 5 (1978-79) Indiana University]

Individual Infants' Discrimination of VOT: Evidence for Three Modes of Voicing

Richard N. Aslin Beth L. Hennessy

David B. Pisoni Alan J. Perey

Indiana University

This paper was presented at the biennial meetings of the Society for Research in Child Development, San Francisco, March 17, 1979. We thank Jerry C. Forshee and David Link for their technical assistance, Diane Kewley-Port for assisting in the speech synthesis, and Wendy Crawford for organizing and implementing the data collection process. This research was partially supported by grants from NIMH (MH-24027-04 and MH-30424-01) and NIH (NS-12179-03 and HD-11915-01).

Individual Infants' Discrimination of VOT: Evidence for Three Modes of Voicing

In the past eight years a great deal of effort has been expended in studying the young infant's ability to perceive voicing differences among stop consonants, particularly stops varying along a voice onset time (VOT) continuum. Much of the controversy surrounding VOT discrimination by infants centers on two issues: the level of analysis involved in VOT discrimination and the role of early language experience. The original findings on infant speech perception (Eimas, Siqueland, Jusczyk and Vigorito, 1971) led many to conclude that infants perceive VOT differences in a linguistically relevant manner; in other words, according to the phonetic categories of the infant's native language. Several more recent lines of evidence, including studies of categorical perception of stop consonants by non-humans (Kuhl and Miller, 1975, 1978) and studies of categorical perception of nonspeech by human adults (Cutting and Rosner, 1974; Miller, Wier, Pastore, Kelly and Dooling, 1976; Pisoni, 1977) and infants (Jusczyk, Rosner, Cutting, Foard and Smith, 1977; Jusczyk, Walley and Pisoni, 1979), strongly suggest that a phonetic level of analysis is not of necessity implicated in the speech discrimination performance of infants. For example, it is now quite clear that categorical perception in and of itself is not sufficient evidence for concluding that a phonetic level of analysis is operative.

Recent evidence from studies of infants whose native language is not English also raise contradictions for a phonetically-based theory of infant VOT perception. Lasky, Syrdal-Lasky and Klein (1975) and Streeter (1976) have shown that infants from Spanish and Kikuyu language environments (respectively) are able to discriminate voicing contrasts which are not phonemic in their native language. And more recently Eilers, Gavin and Wilson (1979) have shown that Spanish infants discriminate VOT contrasts that are phonemic in Spanish and English, but that English infants only discriminate the VOT contrast which is phonemic in English. Of particular interest in these studies is the fact that the voiced - voiceless distinction, cued by VOT values in the plus or voicing lag region, is discriminated by all groups of infants regardless of whether that distinction is phonemic in their native language. Evidence for discrimination of the pre-voiced - voiced distinction, cued by VOT values in the minus or voicing lead region, has only been demonstrated for groups of infants from language environments which employ this distinction phonemically. Thus, if one espoused a phonetically-based theory of infant speech perception, one would be forced to conclude that one phonetic boundary, voiced - voiceless, is primary and universally used by infants regardless of their language environment. Presumably, early experience with the native language can selectively eliminate this primary voiced - voiceless distinction since we know from Lisker and Abramson's (1967) work, illustrated in the first slide, that not all languages use the voiced- voiceless distinction phonemically. This same early experience would presumably create a secondary pre-voiced - voiced distinction provided that this phonetic distinction was used phonemically in the infant's native language.

Insert Slide 1 About Here

Our purpose in conducting the present experiment was to determine whether infants from an English speaking environment, which does not use the pre-voiced - voiced distinction, could in fact discriminate that distinction. Positive evidence among a group of English infants for the ability to discriminate a pre-voiced - voiced distinction would contradict a phonetically-based theory of infant VOT discrimination, if that theory assumed that the pre-voiced - voiced distinction is acquired from early linguistic input.

In order to accurately assess whether infants can discriminate the pre-voiced - voiced distinction we chose a technique which provides individual subject data rather than the more typical between-subject group data employed in most studies of infant speech perception. Our procedure is a modification of the operant head-turning technique developed for use with infants by Eilers, Wilson and Moore (1977). Essentially, the technique capitalizes upon the tendency of infants to orient toward the location of a sound change. As shown in the next slide, an assistant attracts the gaze of the infant who is listening to a repeating speech sound presented from a speaker located on the infant's left.

Insert Slide 2 About Here

The parent and the assistant listen to masking music over headphones so that they are blind as to the timing and nature of experimental conditions. Outside of the sound-attenuated testing booth an experimenter views the infant on a video monitor, and initiates trials on which the repeating background stimulus is changed to a novel or target stimulus. If the infant orients toward the speaker from which the sound change was presented, then the experimenter delivers a brief visual reinforcer which is located adjacent to the speaker. The visual reinforcer consists of an animated toy which remains invisible to the infant until lights within the reinforcement enclosure are turned on by the experimenter.

Our application of this operant head-turning procedure involves three distinct phases: shaping, testing and staircase. All of these phases are controlled by a PDP-11 computer. In the shaping phase a particular VOT stimulus is designated as the background stimulus and is presented once per second. The assistant in the testing booth attracts the infant's gaze approximately 90 degrees to the right of the reinforcer and, provided the infant's head position remains stable, the experimenter outside the booth initiates a trial consisting of three repetitions of the target stimulus. If the infant turns to the left, the experimenter delivers a three second presentation of the visual reinforcer. If the infant fails to respond, the experimenter delivers the reinforcer contingent with the sound change in an attempt to shape the head-turning response. On these initial trials the target stimulus is presented at an intensity level 10 db greater than the level of the background stimulus. The experimenter reduces this intensity cue over the course of shaping until the infant is judged to be responding consistently to the change from the background to the target without the presence of the intensity cue.

At this point the experimenter puts on headphones and the testing phase begins. In this phase there are two types of trials: experimental trials on which the background is changed to the target stimulus and control trials on which the background stimulus is not changed to the target stimulus. During both types of trials the experimenter scores the headturns in an identical manner. The experimenter is unaware of whether a given trial is an experimental or a control trial. A pure tone signal is all that is presented over the experimenter's headphones during the four second scoring interval. If the infant passes our criterion of 80 per cent correct for a minimum of five experimental and five control trials, the computer sends the experiment into the staircase phase.

In the staircase phase the infant's performance on each trial determines the characteristics of the target presented on subsequent trials. In the next slide, we have shown the staircase data from an infant who was shaped and tested with a background stimulus of 0 msec VOT and a target stimulus of -70 msec VOT.

Insert Slide 3 About Here

The staircase follows a 2-up, 1-down algorithm. Two consecutive correct trials are followed by a change in the VOT value of the target which reduces the VOT difference between the target and background. A single incorrect trial results in an increase in the VOT difference between the target and background. Notice that this staircase generates a series of reversal points, and the midpoints of these reversals provide a VOT boundary estimate based on a 70 per cent level of correct responding. Finally, also note that we employed probe trials whenever the infant responded incorrectly on two consecutive trials. These probe trials consisted of presenting the original target stimulus, the one which was most discrepant from the background stimulus. A correct probe trial resulted in a return to the previous target stimulus. Three consecutive incorrect probe trials was our criterion for termination of the experimental session.

A total of 92 infants ranging in age from 5.5 to 11 months participated in our experiment. Fourty-two of these infants failed to show evidence of acquiring the headturn response after 2 sessions and were dropped from the study. Of the remaining 50 infants 44 completed the testing phase on the first of several VOT series; namely a series using either -70 msec VOT as the background and +70 msec as the target or one using +70 as the background and -70 as the target. Twenty-seven of these 44 infants who completed the testing phase went on to complete at least one staircase phase. It should be obvious that this entire procedure is rather rigorous for an infant, typically demanding 50 to 75 trials. Also note that we did not combine data across sessions. Thus, many infants would progress through shaping and testing only to cease responding during the staircase phase before meeting our 5 reversal point criterion.

The stimuli employed in the study were 350 msec in duration and consisted of 5 formant bilabial stops with a following vowel of /a/. The stimuli were modeled after natural speech and were generated on the Klatt (1977) software synthesizer.

The next slide shows the staircase data obtained from the 27 infants who completed one of the two full VOT series. In these full VOT series the staircase stepsize was 20 msec.

Insert Slide 4 About Here

Note that the predominant boundary location regardless of the direction in which the target was shifted centered around the VOT value typically associated with the voiced-voiceless boundary in English. In fact, the mean boundary value from these 27 subjects was 22 msec. On the basis of these data alone one would be likely to conclude that this boundary in the voicing lag or plus region of the VOT continuum is not only primary, but that it is the only one reliably discriminated by English infants. To test this hypothesis further we ran the infants on several truncated VOT series which spanned a more limited range of the -70 to +70 msec continuum.

The next slide shows those subjects who after completing the full VOT series went on to complete the 0 +70 or the +70 0 series. Not surpisingly, these infants also showed a boundary near the English voiced-voiceless boundary, but the smaller step size of 10 msec provides a more accurate estimate of that boundary, which in these data is 30 msec.

Insert Slide 5 About Here

The next slide shows the data from the -20 +50 and +50 -20 series. Again note that the boundary is typically in the voicing lag region, although there are some infants whose boundary falls very close to the -20 msec value.

Insert Slide 6 About Here

Data from the final four VOT series are shown in the next slide.

Insert Slide 7 About Here

First, it is clear that several infants provided data on contrasts in this pre-voiced - voiced region of the VOT continuum, thus providing compelling evidence that these contrasts can be discriminated by English infants. Second, it is clear that many infants did not complete testing on these contrasts. This is in part due to the fact that these were the last VOT series tested on each infant and boredom with the reinforcer accounts for much of our attrition on these contrasts. We also believe, on the basis of a few infants who were run on these contrasts before the other contrasts, that discrimination in the voicing lead region is more difficult that discrimination in the voicing lag region. Nevertheless, the demonstration that some English infants can reliably discriminate across the non-phonemic pre-voiced - voiced boundary raises serious questions concerning a phonetically-based theory of VOT perception in infants, if such a theory proposes that the prevoiced distinction is acquired during early language development.

Finally, one can see in the next slide that those infants who completed several of the VOT series were quite consistent in the location of their category boundaries.

Insert Slide 8 About Here

These data from individual infants is all the more impressive when contrasted with the data from adults who were tested under the same conditions. The last slide shows these adult data and one can see a remarkable similarity in these boundary values compared to the infant data.

Insert Slide 9 About Here

These data could be interpreted as supporting evidence for a phonetically- based theory of VOT perception in infants if one assumed that both lead and lag voicing contrasts are discriminated by all infants regardless of their language environment. However, we believe that these data, along with the data from adult and infant nonspeech studies and cross-species comparisons, provide convincing evidence that a phonetic or interpretive level of analysis is not required to account for the infant VOT discrimination findings. An alternative to a phonetically-based theory of infant VOT perception is to postulate that a basic mechanism of the mammalian auditory system analyzes the relative onset time of complex acoustic signals in a categorical manner. This analysis is characterized by two regions of heightened discriminability along the onset time continuum -- one near +20 msec and one near -20 msec. Moreover, there appears to be a clear bias in favor of the +20 msec region, a bias which offers a reasonable explanation for the failure of several studies to find evidence that infants can discriminate stimuli in the voicing lead region. Whether early experience, either at the acoustic or the phonetic level, can influence an individual infant's sensitivity to a particular region along the onset time continuum remains to be determined in future studies.

References

Cutting, J. E. and Rosner, B. S. Categories and boundaries in speech and music. Perception and Psychophysics, 1974, 16, 564-570.

Eilers, R. E., Gavin, W. J and Wilson, W. R. Linguistic experience and phonemic perception in infancy: A cross-linguistic study. Child Development, in press, 1979.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. and Vigorito, J. Speech perception in infants. Science, 1971, 171, 303-306.

Jusczyk, P. W., Rosner, B. S., Cutting, J. E., Foard, C. F. and Smith, L. B. Categorical perception of non-speech sounds by two-month old infants. Perception and Psychophysics, 1977, 21, 50-54.

Jusczyk, P. W., Walley, A. and Pisoni, D. B. Infants' discrimination of tone onset time differences: Some implications for voicing perception. Paper presented at the biennial meetings of the Society for Research in Child Development, San Francisco, March, 1979.

Klatt, D. H. A cascade/parallel terminal analog speech synthesizer and a strategy for consonant-vowel synthesis. Unpublished manuscript, M.I.T., 1977.

Kuhl, P. K. and Miller, J. D. Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. Science, 1975, 190, 69-72.

Kuhl, P. K. and Miller, J. D. Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. Journal of the Acoustical Society of America, 1978, 63, 905-917.

Lasky, R. E., Syrdal-Lasky, A. and Klein, R. E. VOT discrimination by four to six and a half month old infants from Spanish environments. Journal of Experimental Child Psychology, 1975, 20, 215-225.

Lisker L. and Abramson, A. S. The voicing dimension: Some experiments in comparative phonetics. Proceedings of the 6th International Congress of Phonetic Sciences, Prague, 1967.

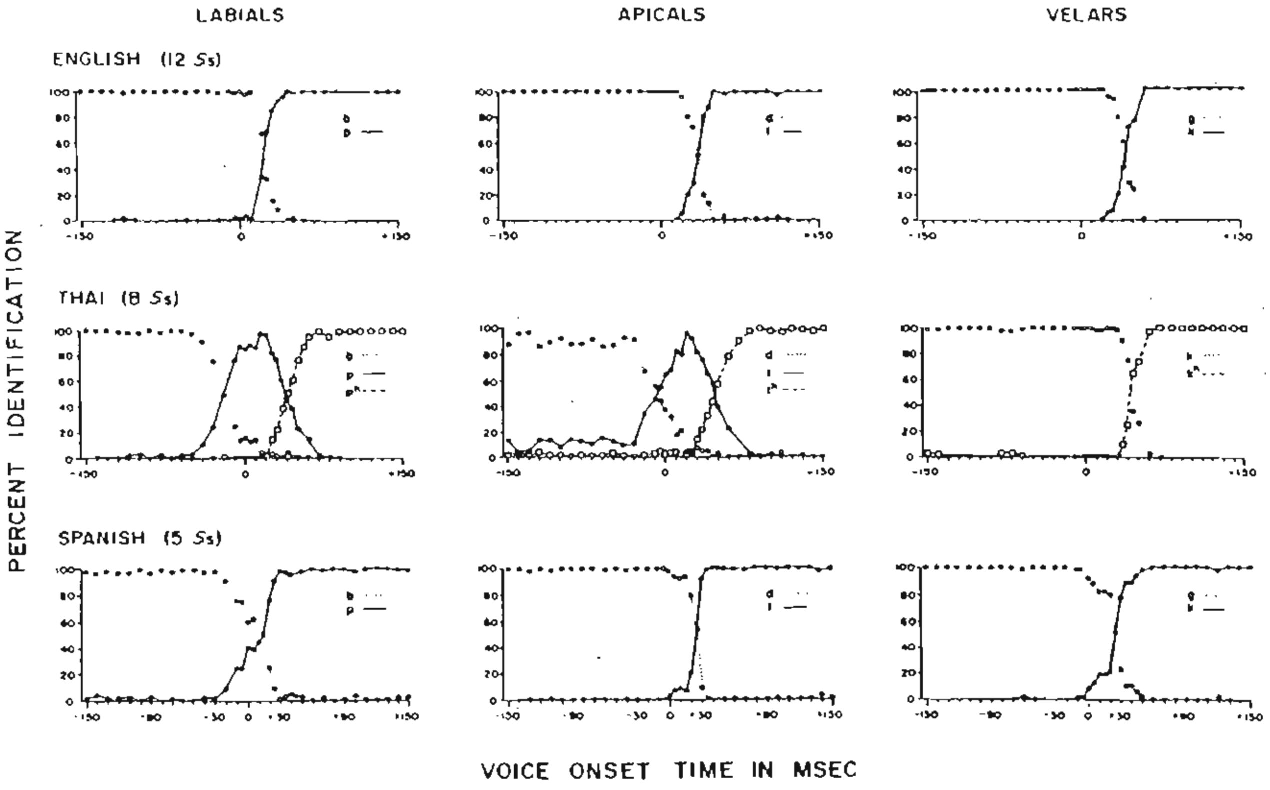
Miller, J. D., Wier, C. C., Pastore, R., Kelly, W. J. and Dooling, R. J. Discrimination and labeling of noise-buzz sequences with varying noise lead times: An example of categorical perception. Journal of the Acoustical Society of America, 1976, 60, 410-417.

Pisoni, D. B. Identification and discrimination of the relative onset of two component tones: Implications for the perception of voicing in stops. Journal of the Acoustical Society of America, 1977, 61, 1352-1361.

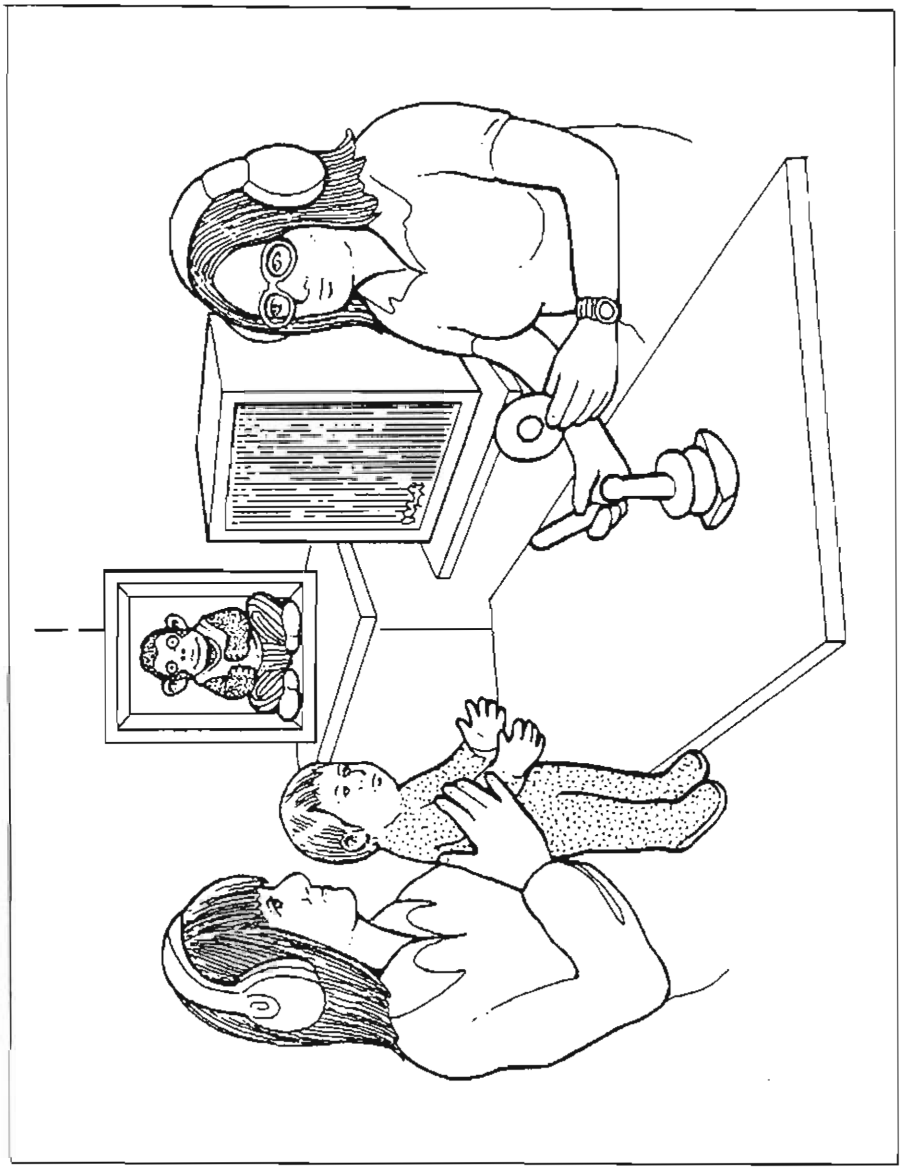
Streeter, L. Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. Nature, 1976, 259, 39-41.

Slide 1

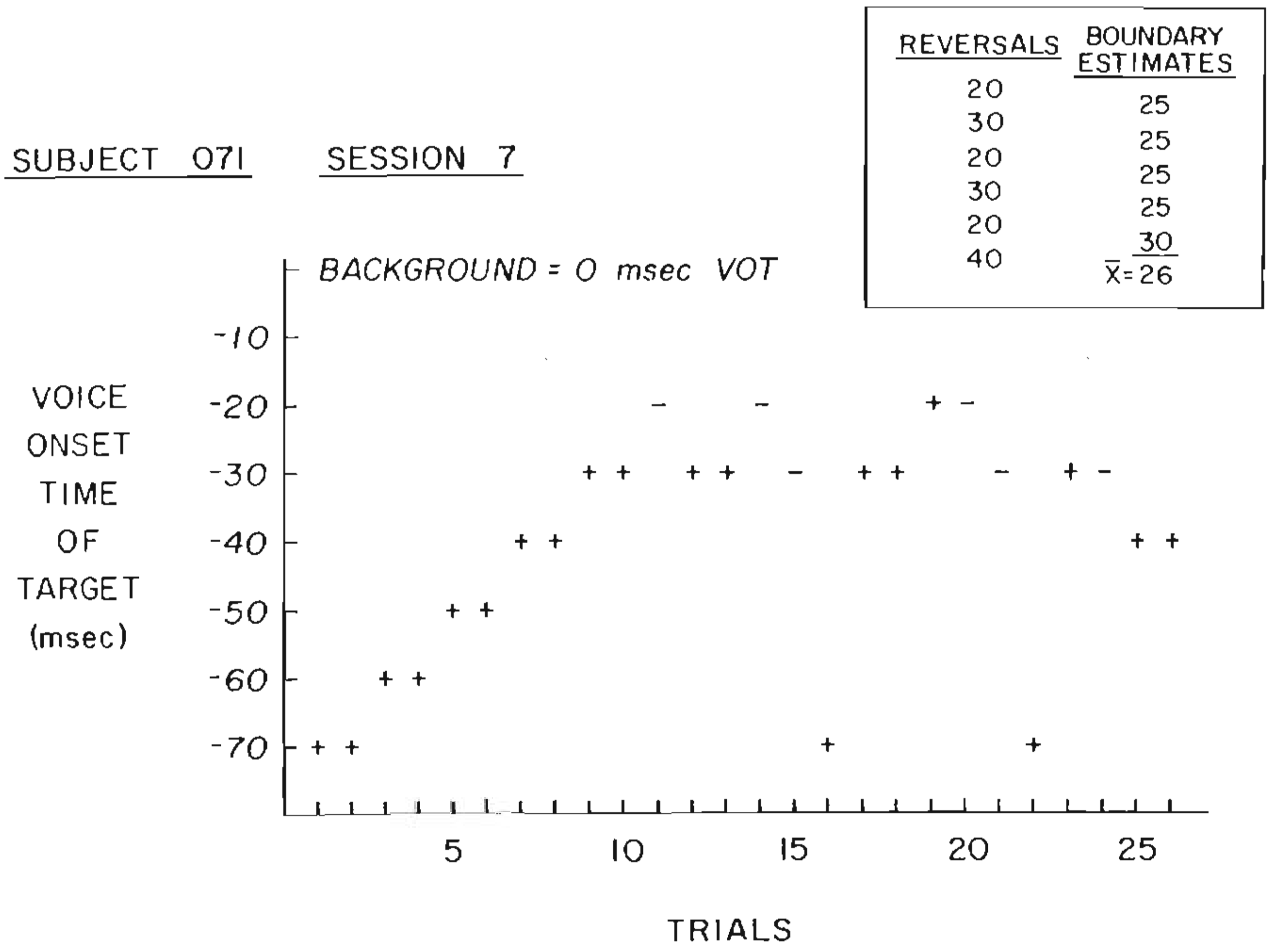
LISKER & ABRAMSON (1967)
CROSS-LANGUAGE LABELING DATA



Slide 2



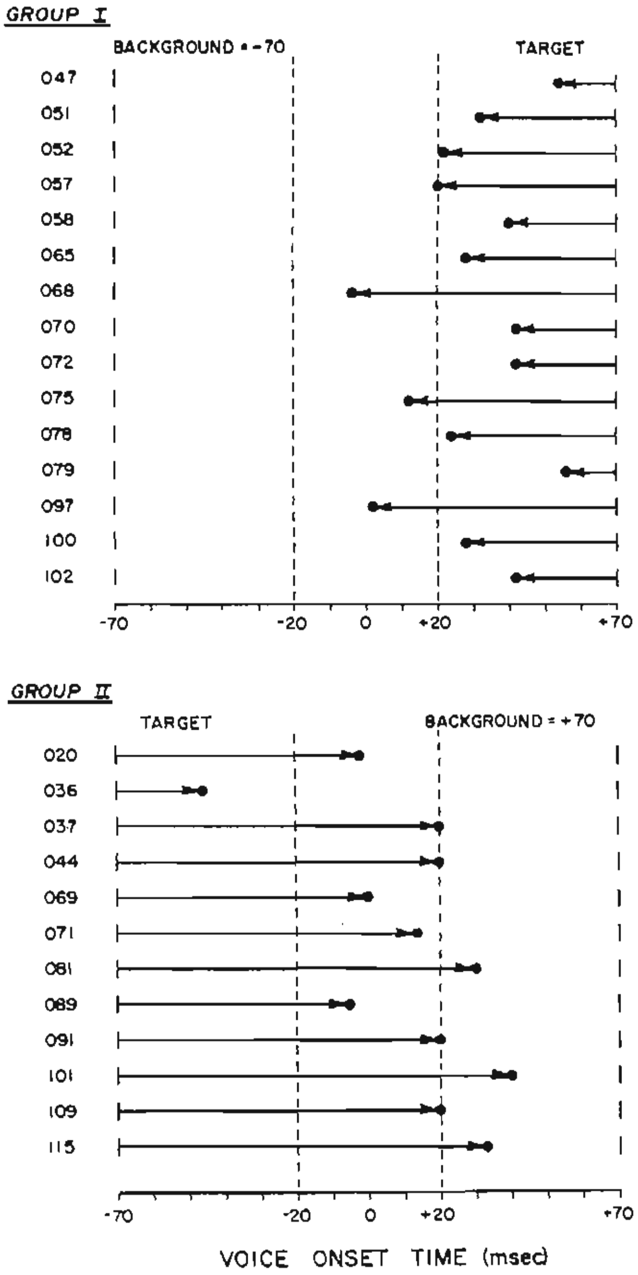
Slide 3



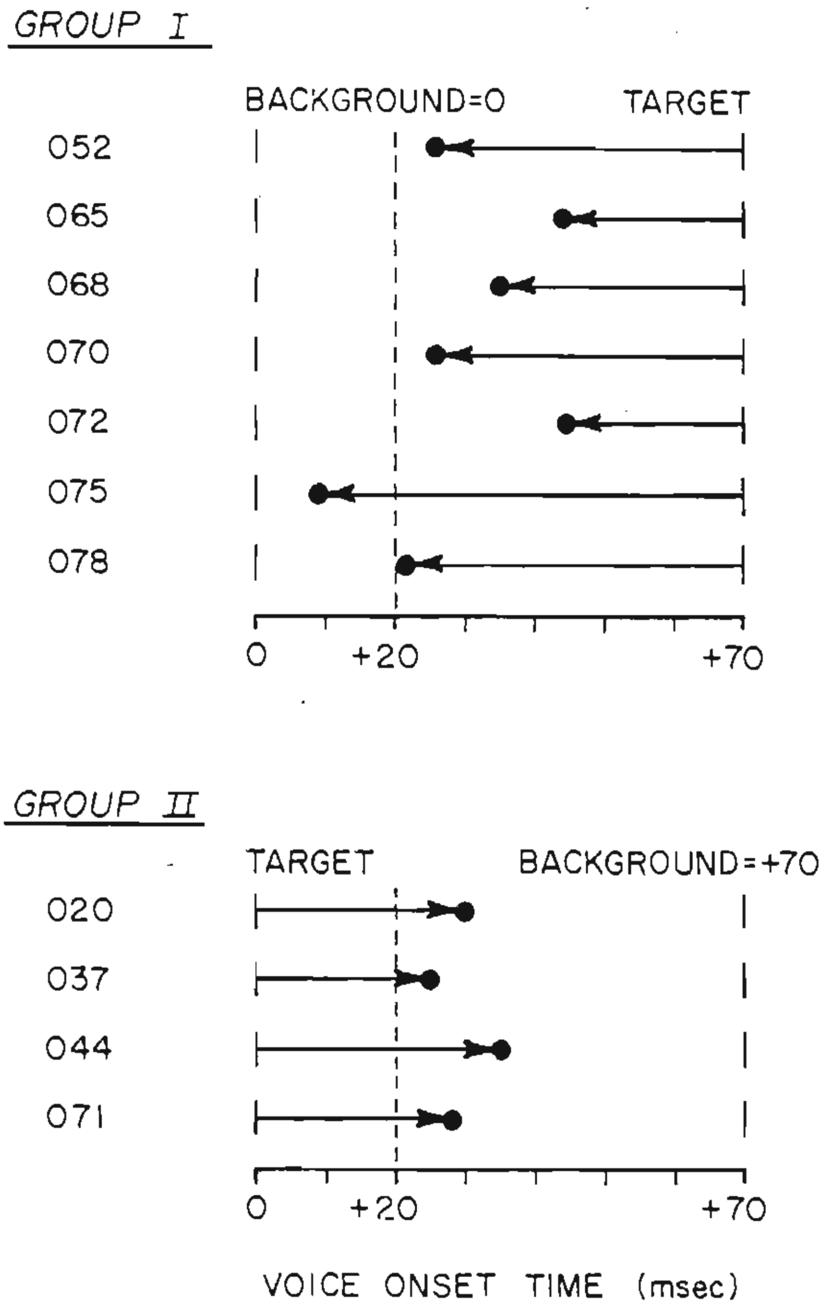
SUBJECT 071 SESSION 7

REVERSALS	BOUNDARY ESTIMATES
20	25
30	25
20	25
30	25
20	30
40	$\bar{X}$=26

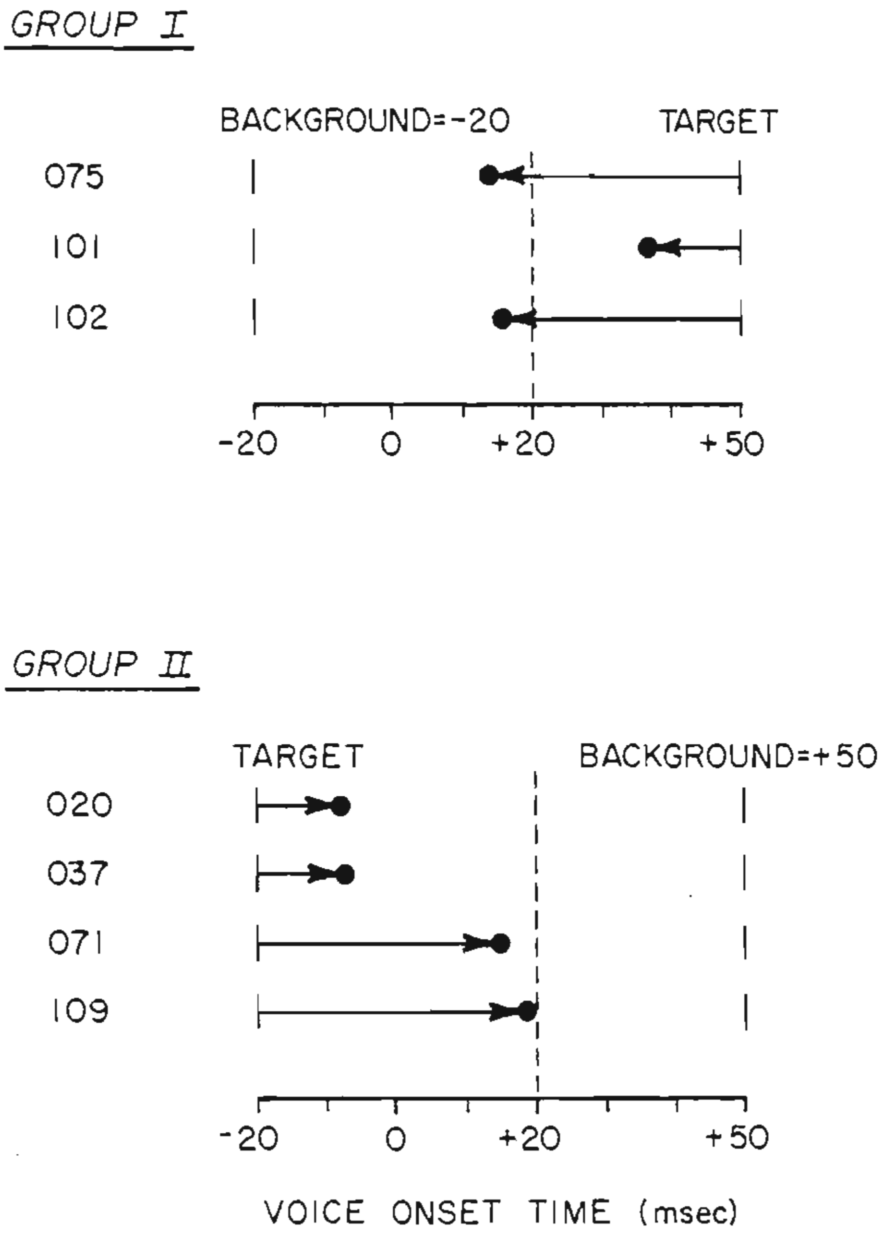
Slide 4


GROUP I
BACKGROUND = -70
TARGET
047
051
052
057
058
065
068
070
072
075
078
079
097
100
102
-70
-20
0
+20
+70
GROUP II
TARGET
BACKGROUND = +70
020
036
037
044
069
071
081
089
091
101
109
115
-70
-20
0
+20
+70
VOICE ONSET TIME (msec)

Slide 5


GROUP I
BACKGROUND=0
TARGET
052
065
068
070
072
075
078
0
+20
+70
GROUP II
TARGET
BACKGROUND=+70
020
037
044
071
0
+20
+70
VOICE ONSET TIME (msec)

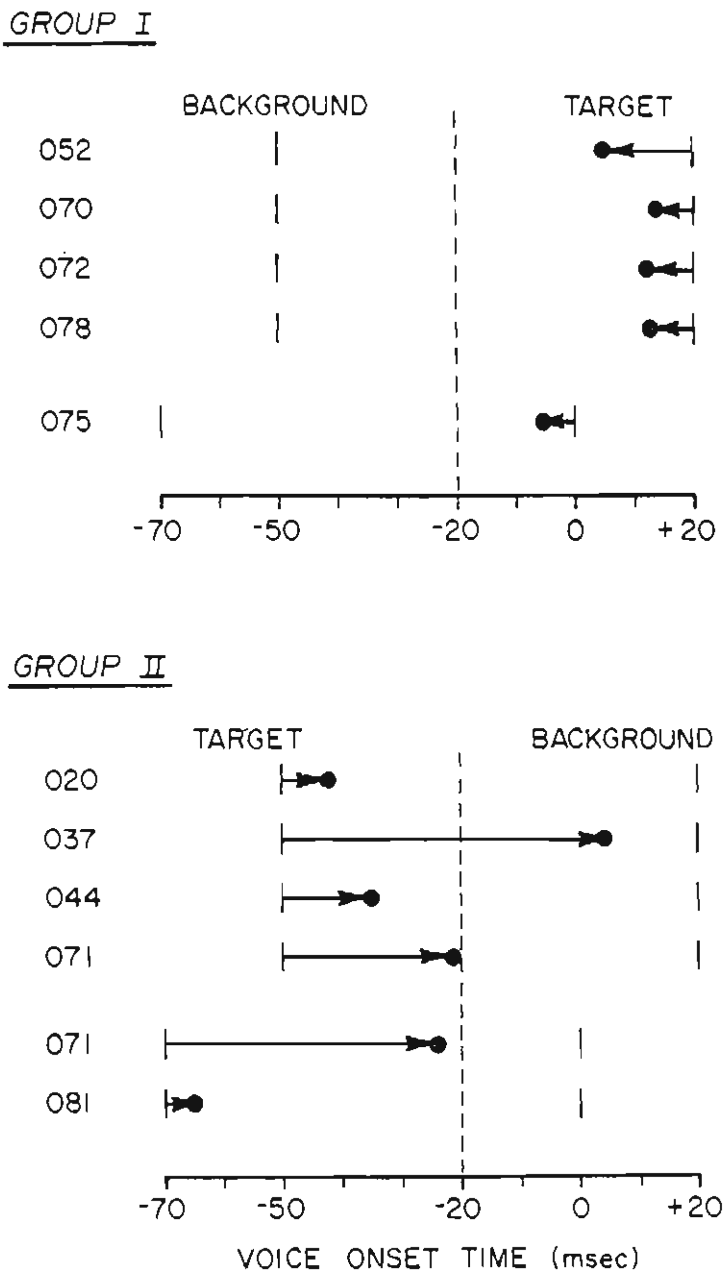
Slide 6



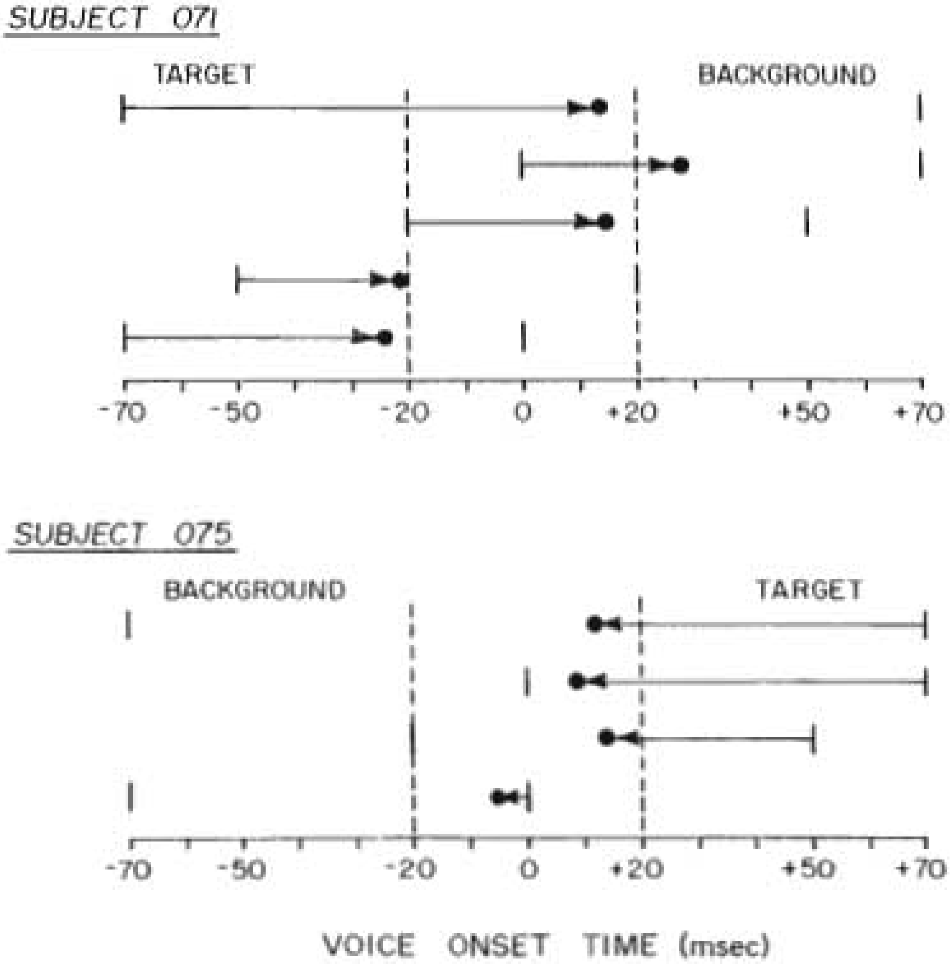
GROUP I

GROUP II

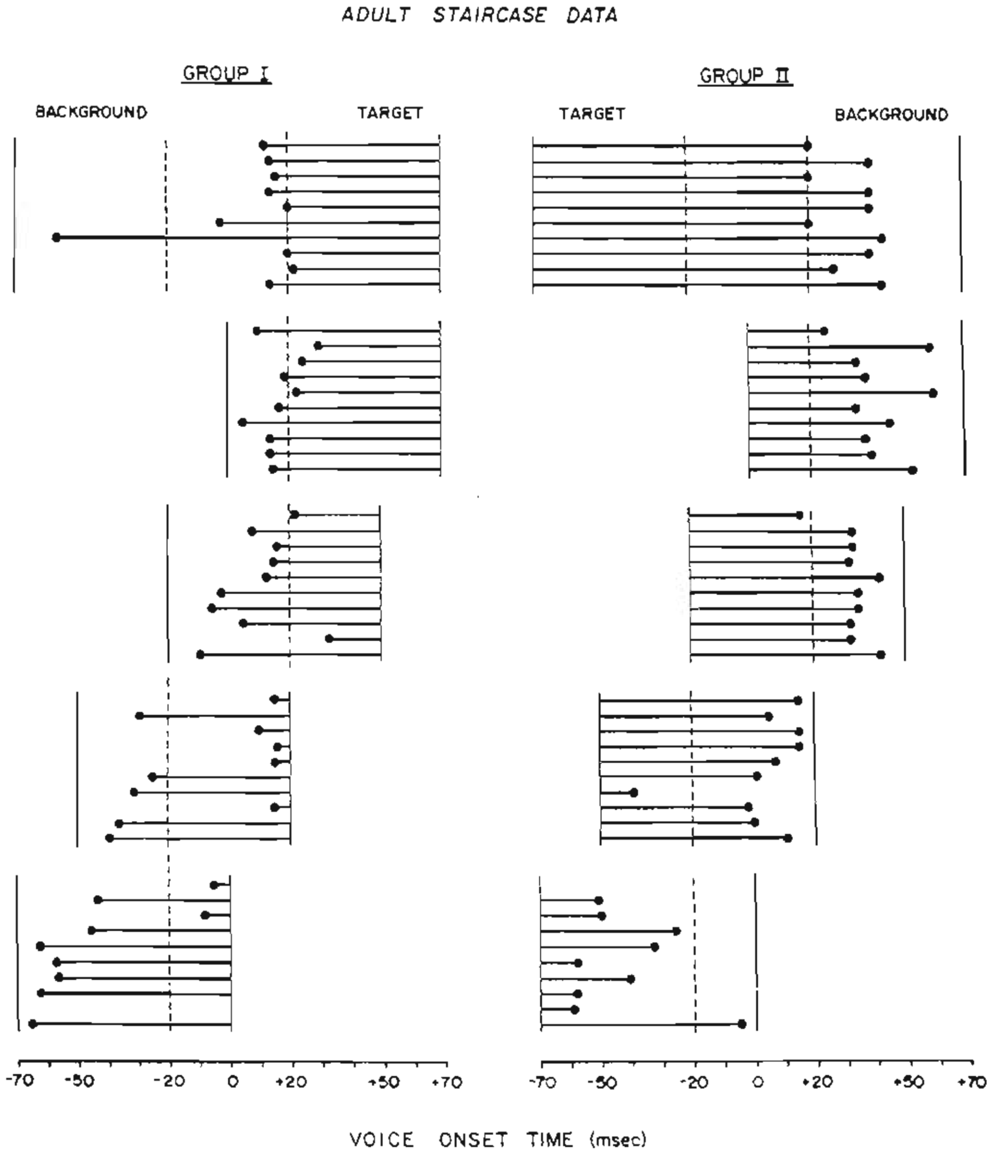
Slide 7


GROUP I
BACKGROUND
TARGET
052
070
072
078
075
-70
-50
-20
0
+20
GROUP II
TARGET
BACKGROUND
020
037
044
071
071
081
-70
-50
-20
0
+20
VOICE ONSET TIME (msec)

Slide 8


SUBJECT 071
TARGET
BACKGROUND
-70
-50
-20
0
+20
+50
+70
SUBJECT 075
BACKGROUND
TARGET
-70
-50
-20
0
+20
+50
+70
VOICE ONSET TIME (msec)

Slide 9


ADULT STAIRCASE DATA
GROUP I
GROUP II
BACKGROUND
TARGET
TARGET
BACKGROUND
-70
-50
-20
0
+20
+50
+70
-70
-50
-20
0
+20
+50
+70
VOICE ONSET TIME (msec)

Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of Eilers, Gavin and Wilson (1979)

Richard N. Aslin and David B. Pisoni

Department of Psychology
Indiana University at Bloomington
Bloomington, Indiana 47405

Short Title: Critique of Eilers, Gavin and Wilson (1979)

Abstract

In a recent report in this journal, Eilers, Gavin and Wilson (1979) presented discrimination data obtained from two groups of infants exposed to different language learning environments. The results showed differences in VOT discrimination between Spanish and English infants suggesting an effect of early linguistic experience. A critique of this study indicates that such conclusions about the effects of early experience on speech perception are unwarranted on both methodological and conceptual grounds. Methodological flaws include the absence of reliable statistical analyses and the failure to guard against experimenter bias effects. Conceptual flaws involve the erroneous interpretation of _failures_ to discriminate certain selected speech contrasts. Inferences concerning the developmental course of speech perception in young infants based on the results of the Eilers et al. study need to be interpreted cautiously in light of these serious criticisms.

Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of Eilers, Gavin and Wilson (1979)[1]

A precise description of the processes and mechanisms used by young infants to perceive the phonetic distinctions of their native language is a difficult challenge that demands careful application of reliable methods of measurement and a reasoned interpretation of empirical findings. Eilers, Gavin and Wilson(1979) recently reported a study of speech discrimination by infants employing a cross-language design. These investigators applied a relatively new and potentially powerful operant head-turning technique to infants selected from two different language learning environments in order to assess the effects of early linguistic experience on the discrimination of voicing in stop consonants. Although the goals and general rationale of this study were quite sound, several serious methodological and interpretative deficiencies led us to question the major conclusions and their implications for the role of early linguistic experience in speech discrimination. The purpose of this report is to point out these problems and also caution other investigators who might follow a similar strategy in using this methodology to study perceptual development in young infants.

Our criticisms of this paper focus on several methodological and interpretative problems including: (a) the number of test trials used to measure discrimination and the statistical analysis of the data, (b)the criterion selected to assess the "state" or degree of "attentiveness" of the infant, (c) the presence of strong experimenter bias effects in collecting the data, and (d) the elimination of potentially important within-subject discrimination data by presenting only grouped means.

The major purpose of the Eilers et al. study was to measure voice onset time (VOT) discrimination for several selected contrasts that are phonologically distinctive in either Spanish or English but not both languages. Based on perceptual data obtained from adult subjects, Eilers et al. selected two critical test pairs for use in their infant study. One stimulus contrast, called the "lag" pair (+10 vs. +40 msec. VOT), was distinctive for adult English subjects whereas the other contrast, the "lead" pair (+10 vs. -20 msec. VOT), was distinctive for adult Spanish subjects. The test stimuli, which differed on VOT, were originally produced on a speech synthesizer at Haskins Laboratories and have been used in earlier investigations of voicing perception in adults and infants (see Lisker and Abramson, 1967; Eimas, Siqueland, Jusczyk and Vigorito, 1971; Lasky, Syrdal-Lasky and Klein, 1975; Streeter, 1976).

In the Eilers et al. study, discrimination of each pair of test stimuli was measured for each infant in an operant head turning procedure (Moore, Thompson & Thompson, 1975; Moore, Wilson & Thompson, 1977; Eilers, Wilson & Moore, 1977). In this procedure the infant is reinforced for a head-turn response toward a loudspeaker when a change occurs in a sequence of repetitive background stimuli. An experimental trial consists of a change from the background stimulus (S1) to a target stimulus (S2). A control trial, on the other hand, consists of no change from the background to target stimulus. A correct response is defined as the presence of a head-turn response on experimental trials and the absence of a head-turn response on control trials. A visual reinforcement is only presented for correct head-turn responses on experimental trials. According to Eilers et al. a measure of discrimination performance can be obtained by simply computing the overall percent correct responses for both experimental and control trials.

In their study, infants received both experimental and control trials for each of the two VOT contrasts, the lag pair and the lead pair. The main hypothesis tested in the study was whether early linguistic experience would differentially affect the likelihood of discriminating the two VOT contrasts. If early linguistic experience is a necessary prerequisite for VOT discrimination, then infants should discriminate only the stimulus pair that is distinctive in their language learning environment. That is, the Spanish infants should discriminate _only_ the lead VOT contrast (i.e., +10 vs. -20 msec.) and the English infants should discriminate _only_ the lag VOT contrast (i.e., +10 vs. +40 msec.). Eilers et al., reported that the English infants discriminated _only_ the lag contrast while the Spanish infants discriminated _both_ the lead and the lag contrasts. They attributed these results to an effect of linguistic experience on speech perception. That is, experience in the language learning environment was presumably responsible for the differences observed in discrimination between the two groups of infants.

The first criticism of this study concerns the number of test trials that were collected for both control and experimental conditions and the subsequent statistical analysis that was carried out on these data. The probability of observing any positive evidence of discrimination is dependent, to a large degree, on the minimum number of test trials employed in the experiment. Measures of discrimination performance, whether based on traditional threshold procedures or more sophisticated methods involving signal detection analysis, require that a sufficient number of trials be collected in order to rule out the possibility that chance factors may have inadvertently affected the observations. In the study reported by Eilers et al., only three experimental and three control trials were collected for each stimulus contrast with each infant tested in the experiment. Estimates of the likelihood of discrimination of either the lag or lead pair were therefore based on only six responses per subject. A criterion of five out of six correct responses was employed in deciding whether each infant discriminated a particular VOT contrast. Unfortunately, this five out or six criterion is not statistically significant at the .05 level according to either the binomial expansion, Fisher's exact test, McNemar's test of correlated proportions, or a X^2 test based on a 2 X 2 contingency table.[2] Six out of six correct responses would be a statistically acceptable criterion according to these procedures but undoubtedly several infants included in the Eilers et al. results did not meet this more stringent criterion. Furthermore, infants who failed to meet the five out of six criterion were re-tested on an "easier" /bit/-/bIt/ contrast. If the infant passed the five out of six criterion on this vowel contrast, then the below-criterion performance on the VOT contrast was accepted as a "real" failure to discriminate. In our opinion, the logic of this procedure is clearly flawed. Not only are some infants providing false evidence of discriminating the VOT contrast and thus inflating their percent correct scores, but other infants are being falsely judged as having discriminated the vowel contrast and thus deflating their low percent correct scores on the VOT contrast. At the very least, infants who *passed* the criterion on the VOT contrast should also have been post-tested on the vowel contrast. Without that post-test, we must conclude that both the evidence for VOT discrimination and the evidence for a failure to show VOT discrimination are inconclusive at best.

The interpretation of negative results is particularly important in a study of this kind because the authors' argument concerning the role of early experience in speech perception rests on establishing that English infants *fail* to discriminate a contrast in voicing that is non-distinctive to adult speakers of

English. Empirical investigations that attempt to "prove" the null hypothesis need extra precautions to insure that their conclusions can be justified on both methodological and statistical grounds. Unfortunately, the results reported by Eilers et al., because they were based on such a small number of test trials do not provide convincing evidence that English infants cannot discriminate a particular voicing contrast. For the same reason, one could also argue that the evidence cited in support of infants' discriminating the other VOT contrasts is also unreliable.

Furthermore, there are additional problems with the particular statistical procedures that were applied to the data. Eilers et al. applied an analysis of variance to their data. But analysis of variance, a parametric statistical test, cannot be used with the data collected in this study because the scores simply do not meet the acceptable criteria required for interval data. With such a small number of trials, the underlying scale could hardly be considered continuous varying over the entire range of values in the scale. The data only meet the criteria for an ordinal scale at best. Moreover, it is highly unlikely that either the normality assumption or the requirement of homogeneity of variances could be met.

A second criticism of this study deals with the procedures used by Eilers et al. to assess the state or attentiveness of an infant during the course of an experimental session. As the authors note, failure to discriminate a difference between two test stimuli, particularly stimulus pairs that may be non-discriminable, could be a "real" effect due to limitations of the infants' perceptual abilities. However, the result could also be due to the attentional state of the infant. To check on an infant's responsiveness in this experiment, Eilers et al. introduced what they believed to be a more easily discriminable contrast, a distinction based on vowel quality between "bit" and "beat". Eilers et al. reasoned that by using a more discriminable contrast such as this, they could determine whether the infant was still "on task" or whether there was a momentary or possibly more lasting change in the state of the infant during the testing session. However, the critical comparisons of interest in this experiment involved whether "fine" discriminative abilities involving the detection of VOT are present in young infants, particularly infants selected from different language-learning environments. Unfortunately, the specific stimulus contrast selected to check on task responsiveness in this study focused the infant's attention on an entirely new and irrelevant stimulus dimension, a dimension involving primarily the detection of spectral (i.e., frequency specific) rather than temporal (i.e., VOT) differences between stimuli. Although there is no universal measure of how well the infant is "on task," one certainly should attempt to direct the infant's attention to the relevant dimension under consideration. The simplest way to accomplish this is to increase the difference in VOT between the background and test stimuli. By introducing a contrast that required the detection of spectral differences between vowels, the infants may have shifted their attention from the particular dimension under investigation (i.e., VOT) to another dimension (e.g., vowel color, stimulus duration, etc.) during the remainder of the test session. Such shifts in attention could easily result in a modification of the infant's criterion for initiating a head-turn response on experimental trials during the course of an experiment, and could well have affected the likelihood of correctly detecting a contrast involving small differences in VOT on subsequent trials. It is interesting to note here that Spanish does not have a contrast between the vowels [i] and [I]. Thus, if early experience does influence the infants' discriminative capacities at this age, a systematic bias would have been introduced by using these particular vowel contrasts to assess attentional state.

Taken together with the earlier criticism concerning the small number of test trials, the procedures used to assess the "state" of the infant could have biased the outcome of the experiment toward finding negative evidence of VOT discrimination in these infants. Both of these criticisms, however, cannot account for the finding that the infants from the Spanish-speaking environment discriminated both of the VOT contrasts, whereas infants from the English-speaking environment discriminated only the lag contrast. This result could simply be due to chance or experimenter bias and may have nothing whatsoever to do with early environmental experience.

Our third criticism of this study relates to the possibility of strong experimenter bias effects operating in the version of the head-turning procedure used by Eilers et al. According to the description provided in the methods section, two independent scorers, one located in the testing booth with the infant (E2), and one controlling the stimulus presentations from outside the booth (E1), recorded whether a head-turn response occurred during each of the six test trials. The authors state that E2 was "blind" to the particular type of trial presented to the infant since this experimenter, as well as the parent, listened to "masking music" over headphones while the experiment was in progress. However, E1, the experimenter located outside the experimental booth, apparently was not blind to the type of stimulus contrast presented on each trial since no in formation was provided in this research report about the procedures taken to insure that experimenter bias was eliminated. In the extreme case, E1 could entirely determine the outcome of a particular trial since agreement between the two scorers was required for a trial to be scored as a correct detection. That is, both scorers were required to respond "yes" on experimental trials for a correct response to be scored and a reinforcement delivered to the infant. If E1 failed to respond on an experimental trial the infant's response would have been scored as incorrect for that particular contrast, indicating that the infant could not discriminate the differences. In the case of control trials, E1 could withhold responding to cancel an incorrect headturn since no reinforcer would be presented if the scorers disagreed. However, the high scorer reliability (i.e., 96%) appears to rule out this potential scorer bias criticism. A more serious scorer bias effect does appear to have been present in the Eilers et al., procedure. Since E1 initiated all trials and therefore knew before each trial began whether an experimental or control trial would be presented, it is possible that E1 could have biased the outcome in the following manner. If E1 initiated an experimental trial when the infant was inattentive to E2 (i.e., the person in the booth playing with toys to attract the infant's gaze), then the infant would be more likely to turn away and consequently that head-turn would be scored as a correct response. Similarly, if E1 initiated a control trial while the infant was highly attentive to E2, then the likelihood of a headturn response on control trials would be reduced substantially. As a result of the procedure that Eilers et al. used, the possiblity of strong scorer bias effects appears quite high, especially since no provision was made, as far as we know, to insure that both scorers were "blind" to the language-learning environment of the individual infants tested or the type of contrast presented on each trial.

The fourth and final methodological criticism of this study concerns the presentation of only group data in each of the conditions of the experiment. The operant head-turning procedure is unique among currently available methods for assessing speech discrimination in young infants because highly reliable within subject data on specific stimulus contrasts can be obtained, provided, of course, that a sufficient number of test trials are obtained from each infant. The absence of individual subject data, taken together with the small number of data points, is especially critical in light of the claims made by these investigators about the potential role that early experience plays in perceptual development. We turn to these considerations in the final section to show how the conclusions of the Eilers et al. study are flawed on both logical and conceptual grounds.

The major interpretive problem with the Eilers et al. study is the claim that "the present results suggest that the Spanish adult lead boundary _is_ precisely the boundary to which infants' perception is tuned through experience with the Spanish language (p. 17)." On close inspection this statement is, however, filled with contradictions and misunderstandings. First, the fact that prelinguistic infants from a Spanish speaking environment also discriminated the VOT contrast that spanned the adult English lag boundary is simply ignored in the discussion of the results. Why should Spanish infants discriminate both VOT contrasts and English infants discriminate only one contrast that was tested in the study? No consideration was given to this rather curious asymmetry in the discrimination data.

Secondly, the authors claim that "linguistic listening experience may be a necessary prerequisite for the acquisition of lead VOT contrasts in infants." But this argument is clearly at variance with the previous cross-language investigations of voicing discrimination reported by Lasky et al. (1975) and Streeter (1976) who found evidence for discrimination of a lead contrast in VOT by both Spanish and Kikuyu infants despite the fact that the specific distinction was non-distinctive in the adult phonological system that the infants were exposed to in the language learning environment. Apparently, Eilers et al. seem to believe that all infants can innately discriminate the lag boundary without any early linguistic experience, but only those infants who actually "hear" the lead contrast used in the postnatal period will acquire the ability to discriminate these VOT contrasts. This conceptualization is especially curious in light of the observations made in the introduction to their paper that "English-learning infants are normally exposed to prevoiced (or lead) stop consonants by English speakers (p.14)..."

How do infants from English-speaking environments learn to selectively attend to precisely those acoustic attributes that are distinctive in the adult phonological system? Eilers et al. propose an explanation for this anomaly by suggesting that infants "must be engaged in an active analysis of the frequency of occurrence of speech sounds at various points along phonetic continua long before these sounds are associated with arbitrary linguistic meanings (p.17)." But if frequency of occurrence were the mechanism responsible for improving discrimination of the lead contrasts, then why do Spanish infants discriminate stimuli in the voicing lag region-- stimuli that are not contrative in the production of stops by Spanish-speaking adults. An alternative account of the Eilers et al. results is possible based on the fact that stimuli in the voicing lead region of the VOT continuum are known to be less discriminable than stimuli in the short lag region.[3] Data supporting the salience of the lag over the lead region is now quite overwhelming (see Lisker and Abramson, 1967; Williams, 1974; Aslin and Pisoni, Note 1; Stevens and Klatt, 1974; Aslin, Pisoni, Hennessy and Perey, 1979). In essence, Eilers et al., have overlooked several serious methodological deficiencies in their attempts to provide support for the null hypothesis that English infants cannot discriminate one particular VOT contrast. Their entire study rests upon the validity and replicability of their failure to show that English infants discriminate the prevoiced-voiced contrast. A single case of an English infant who could discriminate that contrast would offer strong counterevidence for their claim that contrastive use of the voicing lead contrast during early life is a necessary requirement for the acquisition of that discriminative ability. The report of only group means leads one to ask whether some of the infants from the English-speaking environment did, in fact, show evidence of discriminating the lead contrast. In our laboratory, using a similar operant head-turning procedure, but with rigorous controls for experimenter bias effects, we have obtained highly reliable within-subject data indicating that infants from an English-speaking environment can, in fact, discriminate lead contrasts in VOT (see Aslin, Hennessy, Pisoni and Perey, Note 2).

These results, along with our earlier methodological criticisms, indicate that the effects of early experience on speech discrimination have been grossly overestimated by Eilers et al.

In summary, we believe the findings of their cross-linguistic study are seriously flawed on methodological and procedural grounds and that their interpretation of the results is unwarranted. Any inferences or conclusions concerning the developmental course of speech perception in young infants based on the results of the Eilers et al. study should be interpreted cautiously by investigators interested in the effects of early experience on perceptual development.

Reference Notes

Note 1. Aslin, R.N. and Pisoni, D.B. Some developmental processes in speech perception. Paper presented at the NICHD conference "Child Phonology: Perception, Production and Deviation," Bethesda, Maryland, May 28-31, 1978.

Note 2. Aslin, R.N., Hennessy, B., Pisoni, D.B.., and Perey, A.J. Individual infants' discrimination of voice onset time: Evidence for three modes of voicing. Paper presented at the Biennial Meetings of the Society for Research on Child Dvelopment, San Francisco, California, March, 1979.

References

Aslin, R.N., Pisoni, D.B., Hennessy, B. and Perey, A.J. Identification and discrimination of a new linguistic contrast. In J. J. Wolf and D. H. Klatt (Eds.) *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*. New York, N.Y.: Acoustical Society of America, 1979, Pp. 439-442.

Eilers, R.E., Gavin, W. and Wilson; W.R. Linguistic experience and phonemic perception in infancy: A cross-linguistic study. *Child Development*, 1979, *50*, 14-18.

Eilers, R.E., Wilson, W.R. and Moore, J.M. Developmental changes in speech discrimination in infants. *Journal of Speech and Hearing Disorders*, 1977, *20*, 766-780.

Eimas, P.D., Siqueland, E. R., Jusczyk, P. and Vigorito, J. Speech perception in infants. *Science*, 1971, *171*, 303-306.

Hayes, W.L. *Statistics for the Social Sciences, Second Edition*. New York: Holt, Rinehart and Winston, 1973.

Lasky, R.E., Syrdal-Lasky, A. and Klein, R.E. VOT discrimination by four to six and a half month old infants from Spanish environments. *Journal of Experimental Child Psychology*, 1975, *20*, 215-225.

Lisker, L.L. and Abramson, A.S. The voicing dimension: Some experiments in comparative phonetics. *Proceedings of the 6th International Congress of Phonetic Sciences*, Prague, 1967.

Moore, J.M., Thompson, G. and Thompson, M. Auditory localization of infants as a function of reinforcement conditions. *Journal of Speech and Hearing Disorders*, 1975, *40*, 29-34.

Moore, J.M., Wilson, W.R. and Thompson, G. Visual reinforcement of head-turn responses in infants under 12 months of age. *Journal of Speech and Hearing Disorders*, 1977, *42*, 328-334.

Stevens, K.N. and Klatt, D.H. Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 1974, *55*, 653-659.

Streeter, L.A. Language perception of 2 month-old infants shows effects of both innate mechanisms and experience. *Nature*, 1976, *259*, 39-41.

Williams, C.L. Speech perception and production as a function of exposure to a second language. Unpublished Ph.D. thesis, Harvard University, 1974.

Footnotes

[1] The preparation of this paper was supported, in part, by NICHHD grant HD-11915-01, NINCDS grant NS-12179-04 and NIMH grant MH-24027-05 to Indiana University, Bloomington.The paper was written while the second author was a Guggenheim fellow at the Research Laboratory of Electronics, M. I. T. We thank Susan Dumais, Peter Jusczyk, Pat Kuhl and Louis Goldstein for helpful comments on an earlier version of the paper. Address reprint requests to either author in care of the Department of Psychology, Indiana University at Bloomington, Bloomington, IN. 47405.

[2] Hayes (1973) describes each of these four tests (pages 735-742; Table II and pages 830-884). Although a simple X^2 test on 5 out of 6 trials is significant at the .05 level, the low expected cell frequency (<5) in a 2 X 2 table containing only 6 trials demands a correction for continuity. This correction (Hayes, 1973, page 735) results in the failure of 5 out of 6 correct to reach the .05 level of significance. In addition, the trials must also be independent which in this case they are not. According to the binomial expansion tables, a minimal criterion is 8 out of 10 (p=.055), although clearly 15 out of 20 trials would be advisable.

[3] This asymmetry may result from the fact that in stops with voicing lead, the first formant has a relatively low amplitude and the primary cue for discrimination is the duration of the

prevoicing. In contrast, there are several additional and potentially more salient cues for the discrimination of voicing in the lag region, including the presence or absence of an F1 transition, the onset of F1 relative to higher formants and the presence of aspiration noise after the release from stop closure.

FUNDAMENTAL FREQUENCY AS A CUE TO POSTVOCALIC VOICING: SOME DATA FROM PERCEPTION AND PRODUCTION

Thomas M. Gruenenfelder

Department of Psychology

Indiana University

Running head: Fundamental frequency and postvocalic voicing

Abstract

Several studies by Lehiste have reported that changes in fundamental frequency (F0) can serve as a cue to perceived vowel length and, furthermore, the perceived lengthening of the vowel can influence perception of the voicing feature of stop consonants in final position. In Experiment 1, we replicated Lehiste's basic results for stop consonants in final position. Experiment 2 extended these results to postvocalic fricatives. The final consonant of stimuli of intermediate vowel duration was more often perceived as voiced when F0 was falling than when F0 was monotone. In Experiment 3, we examined the F0 contours produced by four talkers before postvocalic stop consonants and fricatives in natural speech for minimal pairs of words differing on voicing. The amount of change of F0 of the vowel was no greater before voiced than voiceless consonants. The present results suggest that the perceptual effect cannot be explained by appealing to regularities in the production of F0 contours on vowels preceeding consonants.

Fundamental Frequency as a Cue to Postvocalic Consonsantal Voicing: Some Data from Perception and Production

Thomas M. Gruenenfelder

Indiana University

Recently, several investigators have noted that perceived vowel duration can be affected by the fundamental frequency (F0) contour of the vowel (e.g. Lehiste, 1976; Pisoni, 1976; Wang, Lehiste, Chang, & Darnovsky, 1976). In particular, vowels with a changing F0 were found to be perceived longer than vowels with a monotone F0. The direction of change seems to have little influence on this effect--vowels with rising F0's and vowels with falling F0's were both perceived as longer than vowels with level F0's (Lehiste, 1976).

Variations in vowel duration have been shown to have at least two important effects on phoneme identification. First, vowel duration systematically influences the identification of the vowel itself (e.g. Stevens, 1959). Two vowels with identical formant frequencies but of different durations are frequently identified as different vowels. This effect is particularly strong for the /ɛ/ (short vowel)-/æ/ (long vowel) distinction. Secondly, vowel duration has been shown to affect the perception of voicing of a following consonant (e.g. Denes, 1954). A consonant following a long vowel is more likely to be perceived as voiced than an acoustically identical consonant following the same vowel of shorter duration.

Vowel duration, then, can serve as a phonemic cue. Because F0 contour affects perceived vowel duration, as has been shown by Lehiste's earlier work, it seems reasonable to suppose that F0 contour can also serve as a phonemic cue. In particular, placing a contour on the F0 of a synthesized vowel should affect its identification in a way similar to lengthening it. Furthermore, consonants following a vowel with a changing F0 should be more likely to be perceived as voiced than consonants following a vowel with a monotone F0. Implicit in both these hypotheses is the notion that the greater either the amount of change of F0, or the rate of change, the greater the effect of the changing F0 should be.

In a recent study, Rosen (1977) has failed to confirm the first of these hypotheses concerning vowel identification. In Swedish, several pairs of vowels are contrasted by duration alone. Using Swedish vowels and Swedish speakers, Rosen failed to find systematic effects of F0 contour on vowel identification--a finding contrary to the hypothesis that F0 contour, through its mediating effect on vowel duration, could serve as a cue to vowel identification.

Lehiste (1977), however, has reported data consistent with the second hypothesis by showing that F0 contour affects the perception of a postvocalic consonant. Lehiste synthesized two stimulus continua, each of which ranged from the word "bat" to the word "bad". The stimuli in each continuum varied on vowel duration. In one continuum, the F0 of the vowel was kept constant at 80 Hz, whereas in the second continuum, the F0 fell from 80 Hz to 60 Hz. These stimuli were presented to listeners who were required to identify them as either "bat" or "bad". The results indicated that perception changed from "bat" to "bad" at a shorter vowel duration for stimuli with a falling F0 than for those with a level F0. Similar results were obtained with a continuum ranging from "beat" to "bead". That is, perception changed from "beat" to "bead" at a shorter vowel duration when F0 was falling than when F0 was level. Lehiste's results support the hypothesis that F0 contour, through its mediating effect on perceived vowel duration, can serve as a cue to the voicing of a postvocalic consonant.

The present paper reports the results of three experiments that examined these effects in more detail. All the experiments were concerned with the use of F0 contour as a cue to postvocalic voicing. The first experiment was designed as a replication of Lehiste's results for stop consonants. Because of Rosen's negative findings for vowel identification, we thought a replication of Lehiste's work would be worthwhile. The second experiment generalized Lehiste's basic finding with stops to postvocalic fricatives. Finally, the third experiment, a production study, examined the F0 contour occurring before voiced and voiceless consonants in natural speech. The production study was aimed at determining whether the perceptual results could be explained by appealing to regularities in the production of F0 contours as a function of postvocalic consonantal voicing.

Experiment 1

Method

Subjects. Four Indiana University students served as subjects. All subjects were right-handed, native speakers of English and were paid $3.00 for participating in a single session, lasting about 75 minutes.

Stimuli. The stimuli for Experiment 1 were constructed using the Klatt (1979) software speech synthesizer implemented on a PDP 11/05 computer in the Speech Perception Laboratory at Indiana University (Kewley-Port, 1978). Two 11-item test continua were synthesized, each ranging from the word "bat" to the word "bad". The parameters of the initial and final consonants were the same for all stimuli. To synthesize the initial /b/, F1 was increased linearly from 575 Hz to 675 Hz, F2 from 1325 Hz to 1425 Hz, and F3 from 2370 Hz to 2470 Hz during the initial 35 msec of the stimulus. The final stop was also synthesized by means of formant transitions. F1 fell linearly from 675 Hz to 280 Hz, F2 rose linearly from 1425 Hz to 1600 Hz, and F3 rose linearly from 2470 Hz to 2930 Hz over the final 40 msec of the stimulus. The final stop was unreleased. The vowel was synthesized using steady-state formants of 675 Hz, 1425 Hz, and 2470 Hz. In addition, F4 and F5 were kept fixed at 3300 Hz and 3700 Hz, respectively, throughout the duration of the stimulus.

In both continua, vowel duration was varied from 100 msec to 300 msec in 20 msec steps. The shortest vowel duration was intended to produce a good exemplar of the word "bat"; the longest a good exemplar of the word "bad". In the Monotone continuum, F0 was held constant at 120 Hz throughout the duration of the stimulus. The Falling continuum was identical to the Monotone continuum, except that the F0 was initially set at 150 Hz and fell linearly to 90 Hz during the second half of the vowel. All stimuli were low-pass filtered at 5000 Hz , stored on a computer disk and later output to subjects in real-time via a 12-bit D-A converter.

Procedure. The stimuli were presented over TDH-39 headphones at a comfortable listening level of about 80 dB SPL. The listeners identified the stimuli in the Falling continuum and Monotone continuum in separate blocks of trials. Two listeners heard the Monotone continuum first and two heard the Falling continuum first. Each of the eleven stimuli in each continuum was presented once in each of twenty blocks. The order of stimulus-presentation was randomized in each block. The listeners were required to identify each stimulus as either "bat" or "bad" by pressing one of two appropriately labeled buttons on a response box interfaced to the computer. Stimuli were separated by a 3 sec interstimulus interval.

Results and Discussion

Figure 1 shows the proportion of "bat" responses as a function of vowel duration for the Monotone continuum (dotted line) and the Falling continuum (solid line). The proportion of "bad" responses shown here is simply 1 minus the proportion of "bat" responses. Each point is based on 20 observations by each of 4 subjects, for a total of 80 observations per point. The general pattern of results is clear. Perception changes from "bat" to "bad" at shorter vowel durations in the Falling continuum than in the Monotone continuum. This results replicates Lehiste's (1977) earlier finding that a changing F0 leads to a higher proportion of voiced responses for postvocalic stops. Three of the four subjects showed a pattern of results consistent with that displayed in Figure 1, while the fourth showed little difference between the two sets of test stimuli.

- - - - - - - - - - - - - - -

Insert Figure 1 about here.

- - - - - - - - - - - - - - -

For each subject, the category boundary for each continuum was computed by linearly interpolating between the two stimuli that scanned the 50% point on the identification function. A t-test for related measures on these boundaries confirmed that perception changed from "bat" to "bad" at shorter vowel durations in the Falling continuum than in the Monotone continuum, $t(3) = 3.51$, $p < .05$.

Thus, the results of Experiment 1 successfully replicated Lehiste's earlier work by showing that changes in F0 contour can influence perception of voicing in postvocalic stops. Experiment 2 was designed to extend these results to fricatives in postvocalic position.

Experiment 2

Method

Subjects. Twelve Indiana University students served as listeners in Experiment 2. All were right-handed, native speakers of English and were paid $3.00 for their participation.

Stimuli. Two 10-item continua, each ranging from the word "cease" to the word "seize", were synthesized on the Klatt synthesizer. For all stimuli, the initial and final fricatives were 155 msec and 170 msec in duration, respectively. The fricatives were synthesized using two noise-bands with center frequencies of 4000 and 4900 Hz, each with a bandwidth of 1000 Hz. In both continua, vowel duration varied from 50 msec to 350 msec in approximately equal logarithmic steps. The shortest duration was intended to produce a good exemplar of "cease"; the longest a good exemplar of "seize". The vowel was synthesized using formant frequencies of 280 Hz, 1850 Hz, 2900 Hz, 3700 Hz, and 4000 Hz. The two continua were identical in all respects except for the F0 contour. In the Monotone continuum, F0 was held constant at 110 Hz throughout the vowel. In the Falling continuum, F0 was initially set at 130 Hz and fell linearly to 90 Hz during the second half of the vowel. These stimuli were low-passed filtered at 5000 Hz, and stored in digital form.

Procedure. The procedure followed that of Experiment 1, except that the "cease-seize" continua were used in place of the "bat-bad" continua. Each of the ten stimuli in each continuum was presented once in each of twenty blocks. Five of the listeners heard the Monotone continuum first, and seven heard the Falling continuum first.

Results and Discussion

The results of Experiment 2 are shown in Figure 2, which plots the proportion of "cease" responses as a function of vowel duration for the Monotone continuum (dotted line) and the Falling continuum (solid line). Each point is based on 20 observations by each of 12 subjects, for a total of 240 observations per point. The results show that perception changes from the voiceless "cease" to the voiced "seize" at shorter vowel durations in the Falling continuum than in the Monotone continuum, suggesting that F0 contour affects perceived vowel duration. Of the 12 subjects, nine showed a pattern of results consistent with that in Figure 2, while two showed no difference between the two continua, and only one showed a reversal of the pattern in Figure 2. Category boundaries were computed as in Experiment 1. Again, statistical analysis confirmed that identification changed from the voiceless "cease" to the voiced "seize" at shorter vowel durations in the Falling continuum than in the Monotone continuum, $t(11) = 5.08$, $p < .001$.

- - - - - - - - - - - - - - -

Insert Figure 2 about here.

- - - - - - - - - - - - - - -

The results of Experiments 1 and 2 provide further support for the hypothesis that F0 contour can serve as a cue for postvocalic voicing, presumably due to its mediating effect on perceived vowel duration. However, since both the present experiments and Lehiste's (1977) earlier study used synthesized speech, they do not address the question of whether the F0 cue is actually deployed systematically in speech production. Experiment 3 was therefore carried out to determine whether systematic differences in F0 contour can be observed in the production of voicing contrasts in final position. The F0 contours occurring before postvocalic consonants were analyzed for minimal pairs of words differing on the voicing of the final consonant.

Experiment 3

Method

Subjects. Four laboratory personnel volunteered as subjects. All were male native-speakers of English. Three had had some phonetic training.

Materials. The basic set of materials consisted of a set of 23 minimal pairs of words, differing on the voicing of the final consonant. Sixteen of the pairs contained a final stop and seven pairs contained a final fricative. Several different lists were constructed by randomly intermixing these 46 words with 46 additional filler words. The 46 stimulus words are shown in Table 1.

- - - - - - - - - - - - - -

Insert Table 1 about here.

- - - - - - - - - - - - - -

Procedure. Each speaker read through the word list once in citation form in a sound-attenuated room. Each pronunciation was recorded on audio tape and then digitized via a 12-bit A-D converter for analysis. The F0 contours of JS and LG were analyzed using a computer algorithm implemented at the Research Laboratory of Electronics at M.I.T., and similar to the algorithm described in Gold and Rabiner (1969). In this algorithm, the speech signal is initially low-passed filtered to eliminate frequencies above the first formant region. The algorithm then determines the heights and positions of positive and negative peaks, as well as the difference in heights from peak to valley (negative peak), valley to peak, peak to previos peak, and valley to previous valley. Each of these measures is used to arrive at an independent measure of F0 and a complex decision rule is used to determine a final "best" measure of F0 for any point in time. An autocorrelation routine, implemented in the Speech Perception Laboratory at Indiana University, was used to analyze the F0 contours of speakers AW and TC. An analysis of the productions of JS using the autocorrelation routine showed no significant differences from the analysis using the M.I.T. algorithm.

Results and Discussion

Table 2 shows the mean vowel duration (D), F0 at the start of the vowel (F0s), F0 at the end of the vowel (F0e), amount of change of F0 (ΔF0 = F0s-F0e), and the rate of change of F0 (Hz/msec) for final fricatives and final stops for each of the four speakers. The pair "bowed-boat" was not analyzed for speakers LG and TC since since they pronounced "bowed" as /baud/ rather than /bod/. As expected, vowel duration was longer before voiced than voiceless consonants. This held true for all 23 minimal pairs for each speaker.

The results of most interest concern the amount of change of F0 before voiced and voiceless final consonants. The F0 contour of speaker LG fell 9 Hz more before a final voiced stop than before a voiceless stop, t(14) = 3.78, p < .01. Similarly, speaker AW showed a greater drop in F0 before voiced than voiceless fricatives, t(6) = 3.76, p < .01. These results are consistent with the hypothesis that F0 contour is produced by talkers as a cue to final consonant voicing. However, the remainder of the production data are not consistent with that hypothesis. Speaker LG showed no difference in the amount of change of F0 before voiced and voiceless fricatives. The same is true for speaker AW's fricative data, and both the fricative and stop data for speakers JS and TC. These negative results are particularly striking given the much longer vowel durations before voiced than voiceless consonants. In earlier studies, Lea (1973) and Mohr (1971) also failed to observe a systematic influence of the voicing of a postvocalic consonant on the F0 contour of the preceding vowel. The primary difference between the present study and these earlier studies is that the present study analyzed only English words, whereas the earlier studies included a number of nonsense words as well.

- - - - - - - - - - - - - - -

Insert Table 2 about here.

- - - - - - - - - - - - - - -

Taken together, the results of the present set of experiments produce a mixed picture of the relationship between F0 and voicing of postvocalic consonants. While the perceptual data from the present experiments clearly show a consistent effect of variations in F0 contour on voicing in perception of both stops and fricatives, the same consistent relationship was not observed in the analysis of F0 in speech production. Moreover, if we examine

F0 and vowel duration together, our results indicate that rate of change of F0 is less before voiced than voiceless consonants, a result that is precisely the reverse of what had been observed in the perceptual studies. Thus, F0 contour is capable of cueing the voicing of a postvocalic consonant. However, the generality of the F0 contour as a perceptual cue must be qualified since it is not reliably produced by all talkers in the same context.

Work by other investigators (e.g. Haggard, Ambler, & Callow, 1970) has shown that the F0 contour of the initial portion of a consonant in a CV syllable is sufficient to cue the voicing of an initial consonant. An F0 contour which is high and falling during the early portion of the vowel indicates that the preceding consonant was voiceless, whereas a low and rising F0 indicates a voiced consonant. Studies of speech production (Mohr, 1971) have indicated that the differing F0 contours associated with voiced and voiceless consonants are due to physiological mechanisms involved in articulation. That the perceptual system knows something about such a physiologically- determined cue is not surprising. However, Experiment 3 of the present paper showed that the F0 contour cue to postvocalic voicing is not reliably produced, and consequently the perceptual results cannot be due to mediation by the articulatory system. Consequently, any account of the use of F0 as a cue to postvocalic voicing cannot have recourse to systematic regularities in speech production.

In summary, F0 contour can serve as a perceptual cue to the voicing of postvocalic consonants. However, this cue is not reliably produced by speakers, indicating that its ability to serve as a perceptual cue is not tied to its regularity in production.

References

Denes, P. Effects of duration on the perception of voicing. *Journal of the Acoustical Society of America*, 1954, *27*, 761-764.

Gold, B., & Rabiner, L. Parallel processing techniques for estimating pitch periods of speech in the time domain. *Journal of the Acoustical Society of America*, 1969, *46*, 442-448.

Haggard, M., Ambler, S., & Callow, M. Pitch as a voicing cue. *Journal of the Acoustical Society of America*, 1970, *47*, 613-617.

Kewley-Port, D. KLTEXC: Executive program to implement the KLATT software speech synthesizer. *Research on Speech Perception: Progress Report No. 4*. Department of Psychology, Indiana University, Bloomington, Indiana, 1978.

Klatt, D. H. Software for a cascade/parallel speech synthesizer. *Journal of the Acoustical Society of America*, 1979, in press.

Lea, W. A. Segmental and suprasegmental influences on fundamental frequency contours. In L. M. Hyman (Ed.), *Consonant types and tones*. Southern California Occasional Papers in Linguistics, 1973.

Lehiste, I. Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics*, 1976, *4*, 113,117.

Lehiste, I. Contribution of pitch to the perception of segmental quality. Paper presented at the 9th International Congress on Acoustics. Madrid, 1977.

Mohr, B. Intrinsic variations in the speech signal. Phonetica, 1971, 23, 65-93.

Pisoni, D. B. Fundamental frequency and perceived vowel duration. Paper presented at the 91st meeting of the Acoustical Society of America, Washington, D. C., April, 1976.

Rosen, S. M. Fundamental frequency patterns and the long-short vowel distinction in Swedish. STL-QPSR 1/1977, 31-37.

Stevens, K. N. Effect of duration upon vowel identification. Journal of the Acoustical Society of America, 1959, 31, 109 (A).

Wang, Wm. S.-Y., Lehiste, I., Chang, C.-K., & Darnovsky, M. Perception of vowel duration. Paper presented at the 92nd meeting of the Acoustical Society of America, San Diego, November, 1976.

Table 1

Stimulus Words Used in Experiment 3

Stops

Voiced	Voiceless
Bad	Bat
Bead	Beat
Bid	Bit
Bowed	Boat
Bud	But
Cued	Cute
Jog	Jock
Lube	Loop
Need	Neat
Pad	Pat
Pod	Pot
Rib	Rude
Rude	Root
Seed	Seat
Sued	Suit
Weed	Wheat

Fricatives

Voiced	Voiceless
Buzz	Bus
Eyes	Ice
Maize	Mace
Peas	Piece
Plays	Place
Rise	Rice
Seize	Cease

Table 2

Vowel Duration (D), Starting F0 (F0s), Ending F0 (F0e), Change of F0 (F0), and Rate of Change of F0 (Hz/msec) as a Function of Postvocalic Consonant Voicing

	D	F0s	F0e	F0	Hz/msec
Speaker: JS					
Fricatives					
Voiced	256	95	76	19	.086
Voiceless	142	98	78	20	.151
Stops					
Voiced	200	96	79	17	.084
Voiceless	115	98	80	18	.167
Speaker: LG					
Fricatives					
Voiced	189	120	87	33	.182
Voiceless	91	120	89	31	.353
Stops					
Voiced	200	124	92	32	.166
Voiceless	89	119	96	23	.287
Speaker: TC					
Fricatives					
Voiced	278	132	114	18	.119
Voiceless	143	137	116	21	.260
Stops					
Voiced	230	156	132	24	.110
Voiceless	113	160	140	20	.192
Speaker: AW					
Fricatives					
Voiced	237	117	86	31	.140
Voiceless	119	122	100	22	.198
Stops					
Voiced	208	115	88	27	.135
Voiceless	114	122	100	22	.200

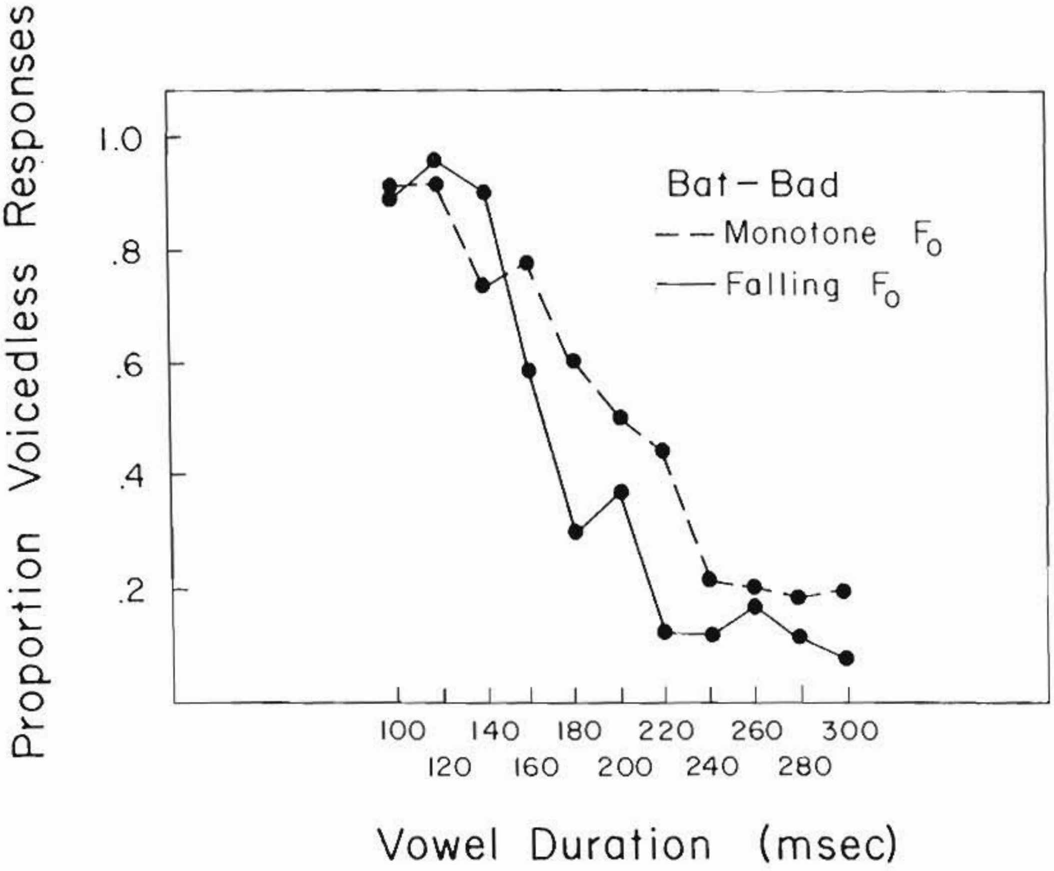


Figure 1. Proportion "bat" responses as a function of vowel duration.

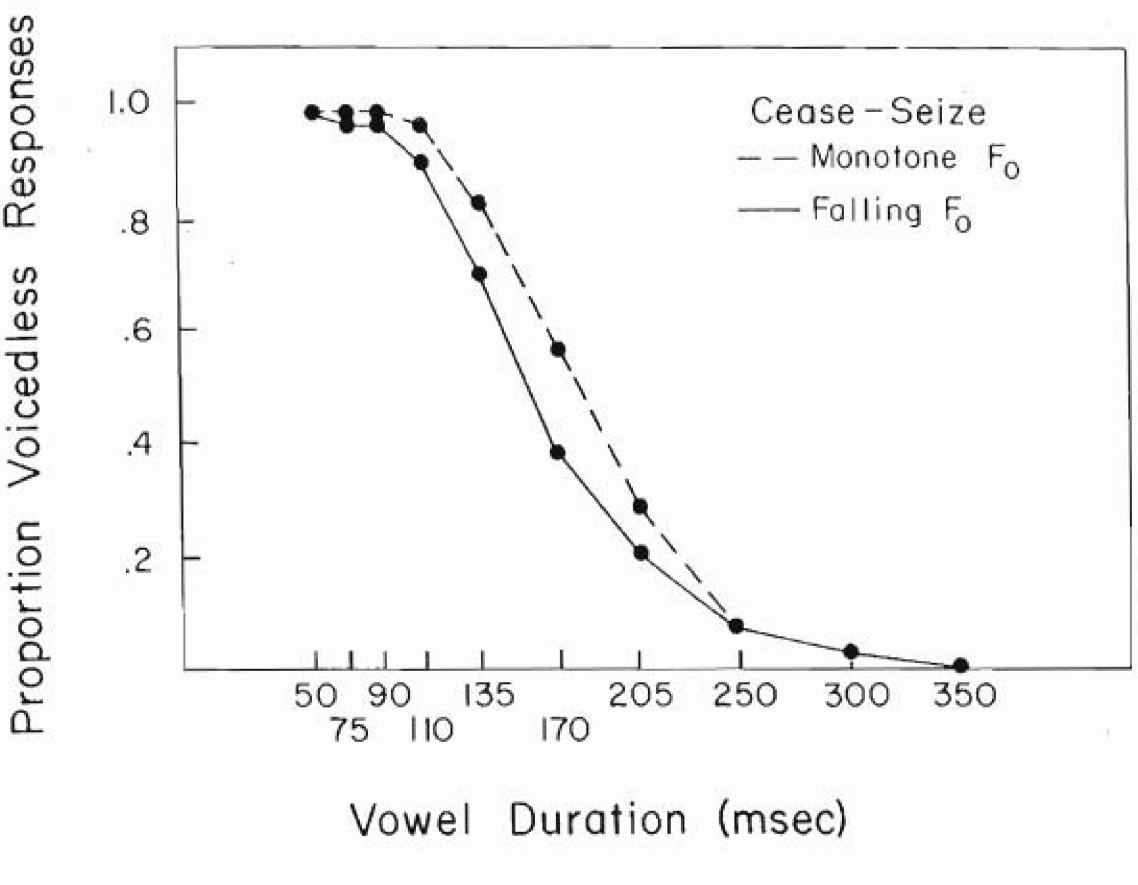


Figure 2. Proportion "cease" responses as a function of vowel duration.

Measuring Lexical Access during Sentence Processing

Michelle A. Blank
Department of Psychology
Indiana University
Bloomington, Indiana 47405

Short Title: Measuring Lexical Access

Abstract

The results from "on-line" investigations of sentence comprehension are often difficult to interpret since it is not always apparent which component processes are reflected in the response measure. The results of two experiments reported here indicate that Phoneme Triggered Lexical Decision (PTLD) response latencies reflect the time needed for lexical access during sentence processing. Listeners were presented with sentences and were asked to make a word/nonword judgment for items beginning with a particular word-initial target phoneme. Speed of lexical access was manipulated by varying the semantic predictability of the target-bearing word. WORD judgments were faster for words that were preceded by semantically related verbs than WORD judgments for words that were preceded by neutral verbs. The present results are consistent with other studies showing semantic facilitation of lexical access during the processing of fluent speech. It is argued that the phoneme-triggered- lexical-decision task is a more suitable measure of lexical access during sentence processes than phoneme monitoring (Foss, 1969) or word monitoring (Marslen-Wilson & Tyler, 1975). In addition, it is pointed out that the phoneme-triggered-lexical-decision task lends itself to modifications which should enable investigators to study various aspects of on-line sentence processing.

Measuring Lexical Access during Sentence Processing*

The problem of discovering how listeners understand spoken language is unquestionably a formidable one. A first step taken by psycholinguists towards solving this problem has been to view sentence comprehension in terms of several component processes. Presumably, these processes include at least an acoustic analysis of the incoming waveform, phonetic and phonological analyses, lexical look-up, as well as syntactic and semantic integration (see, e.g., Studdert-Kennedy, 1974). The results from a large number of studies suggest that these processes combine in such a way as to form a "dynamic-interactive" system (e.g., Foss & Blank, in press; Marslen-Wilson, 1976; Marslen-Wilson and Welsh, 1978). In a system of this kind, computations being executed by each process are influenced by other on-going analyses occurring at several processing levels.

A variety of tasks have been used to study the component processes of sentence comprehension. It would be unrealistic and beyond the scope of this paper to review each of these techniques. Instead, it suffices to say that they can be classified into one of two groups: those that reflect post-sentence comprehension processes, and those that reflect on-line processing. Because of the speed and apparent interactive nature of the subprocesses involved in understanding sentences, it appears that on-line measurements obtained during sentence comprehension are valuable tools for investigating sentence processing. A few examples of tasks which have been assumed to reflect on-going sentence processing include speech shadowing (e.g., Marslen-Wilson, 1973), phoneme-monitoring (e.g., Foss, 1969), word-monitoring (e.g., Marslen-Wilson, 1975) and mispronunciation detection (e.g., Cole, 1973). It should be noted, however, that for many on-line techniques a complete specification of the perceptual processes involved in performing the task is not available. This is in no way a trivial point given that task demands are known to affect the perceptual organization and encoding of stimuli (e.g., Ammon, Ostrowski and Alward, 1971; Aaronson, 1976; Cary, 1971). Thus, psycholinguists need, not only tasks that measure immediate processing, but also knowledge about how such tasks interact with sentence processing mechanisms. Without an explicit set of assumptions about these interactions it is difficult to accurately interpret reaction time data obtained using any of these techinques (See Foss & Blank, in press, for further discussion of this point).

The purpose of this report is to introduce a new task which will hopefully be a useful addition to the psycholinguist's experimental arsenal of on-line measures of sentence processing. Moreover, an attempt will be made to specify at least some of the underlying processes involved in the task.

Lexical decision and phoneme-monitoring are two tasks which have been used in numerous psycholinguistic studies (e.g., Blank & Foss, 1978; Forster & Chambers, 1973; Foss, 1970; Foss & Jenkins, 1973; Meyer, Schvaneveldt & Ruddy, 1975; Scarborough, Cortese & Scarborough, 1977). The new task, phoneme-triggered-lexical-decision (PTLD), is the result of combining these two techniques. Before describing PTLD, a brief description of lexical decision and phoneme monitoring is needed.

In lexical decision experiments subjects are presented with a series of individual words and nonwords and are asked to make a word/nonword judgment for each item. Response times are measured from the onset of the item to the execution of the response. Latencies are taken as a measure of lexical access. In phoneme-monitoring, subjects listen to a list of individual sentences. Each sentence is preceded by the specification of a target phoneme. The subjects' task is to press a button when they hear the word-initial target phoneme specified for that sentence. Reaction times obtained using this task represent the time elapsed between the onset of the target phoneme and the subjects' response. These times are taken to reflect the relative complexity of comprehension, or, more germane to the present discussion, of one or more of its component processes.

In the PTLD task, listeners are asked to make a word/nonword judgment for a particular word occurring within a sentence. The word which listeners judge is the one which begins with the target sound specified for that sentence. Listeners indicate their response by pressing one of two buttons labelled either "real word" or "nonsense word." Table 1 lists sample sentences which I have used in PTLD studies. Each sentence is preceded with the specified target phoneme for that sentence and the correct response is indicated for the underlined target word. In short, listeners use the target phoneme to "trigger" their lexical decisions (i.e., their word/nonword judgments). Like phoneme-monitoring, response times are measured from the onset of the target phoneme to the time one of the buttons is depressed. The pilot study and main experiment which are reported here bear on the question of whether or not phoneme-triggered-lexical-decisions can be used to investigate lexical access as it occurs during sentence comprehension.

Insert Table 1 about here

At this point, a word is in order about the relationship between phoneme-triggered-lexical-decision and phoneme monitoring. Phoneme monitoring has, in the past, been used with some success to investigate "on-line" lexical processing during sentence comprehension (e.g., Blank & Foss, 1978; Foss & Jenkins, 1973). What is the important difference between phoneme monitoring and

Table 1

Example sentences used in

Phoneme Triggered Lexical Decision studies
(P.T.L.D.)

Trigger phoneme	Sentence	Correct Response
/p/	Before going to bed, the traveling salesman packed his suitcase.	WORD
/t/	The famous surgeon skillfully removed the tadgy bullet fragments from the victims chest.	NONWORD
/d/	Within a matter of moments, the tacreela destroyed all of the homes along this side of the street.	WORD

phoneme-triggered-lexical-decisions? One of the assumptions underlying the interpretation of phoneme monitoring response times is that listeners respond to target phonemes by referencing an abstract sound representation of the target-bearing word. This information about a word's sound pattern is presumably stored in the mental lexicon and becomes available subsequent to word identification (Foss & Swinney, 1973; Marslen-Wilson & Tyler, Note 1; Morton & Long, 1976). A corollary assumption is that phoneme monitoring latencies reflect the time needed to access the target-bearing word. However, recent evidence suggests that listeners can use stimulus (that is, acoustic/phonetic) information to identify phonemes (Foss & Blank, in press; Newman & Dell, 1978). For word-initial phonemes, this "sensory-based" information is derived independent of, and prior to, lexical access. Thus, it appears that phoneme monitoring response times may not always reflect lexical access of the target-bearing word. Whenever a target phoneme is identified on the basis of stimulus information, a response could be initiated before lexical access. The notion that subjects in phoneme monitoring experiments "can" or "might" identify phonemes on the basis of sensory information is not a new one (see, Foss & Lynch, 1969). However, this view was subsequently abandoned in favor of the assumption that a phoneme is identified "on the basis of the internal response following recognition of the word containing it," (Morton & Long, 1976, p. 43). Now that questions have been raised concerning the perceptual processes underlying phoneme-monitoring, the interpretations of some earlier monitoring studies must be reconsidered.

In contrast to phoneme monitoring, the task demands of phoneme-triggered-lexical-decision insure that responses will be executed after the target-bearing word has been accessed (i.e., post lexically). Although phoneme monitoring is one component of PTLD, it is a necessary, not sufficient process for an appropriate response to be made. A lexical decision can never be made without consulting the lexicon. This is because listeners cannot know whether a given stimulus input corresponds to a real word or a nonsense word until lexical access is attempted. Only after lexical access can listeners determine if a given acoustic signal has a corresponding abstract sound representation stored in the lexicon with semantic and syntactic information associated with it. Thus, unlike phoneme-monitoring, phoneme-triggered-lexical-decision guarantees that lexical access has occurred prior to execution of the response. Consequently, PTLD latencies should provide on-line measures of lexical access.

Pilot Study

A pilot study was conducted to simply test whether subjects could perform this task with relative ease and no adverse effects on comprehension. The materials used in this experiment were originally designed for a phoneme monitoring study reported by

Foss & Blank (in press). However, by merely changing the instructions these materials could be used to obtain information relevant to the present issue.

Eighteen experimental sentences were varied in the following manner: the target-bearing item was either a real word or a nonsense word. An example sentence with /p/ as the triggering phoneme is given below:

> The inquisitive young investigator annoyed the (prominent/pradament) businessman on trial for misuse of funds.

Each subject heard nine real word sentences and nine nonsense word sentences; no subject heard both versions of the same experimental sentence. Eight subjects heard a total of 72 sentences (18 experimental/54 filler). Handedness was counterbalanced across word/nonword button positions. Subjects were forwarned in the instructions that after they heard all the sentences a comprehension test would follow. This instruction was intended to emphasize the importance of paying close attention to the meaning of the sentences. The comprehension test was actually a recognition task consisting of twenty-four sentences. Half of these sentences were old, the subjects had heard them during the experiment, and half were new. Subjects were instructed to state whether each sentence was old or new. All of the old sentences were chosen from among those fillers which contained only real English words.

The results of this pilot study were as follows: mean reaction time for word targets was 857 milliseconds, whereas mean reaction time for nonsense word targets was 1111 milliseconds. These response times are in accord with a well-documented effect obtained in lexical decisions studies -- longer latencies for nonwords as compared to words. This difference was significant by a sign test, $p < .01$.

Since this study was not designed to directly test whether observed RTs reflect the process of searching the mental lexicon for an entry, no further statistical analysis were done on these data. The main problem with using the observed RT difference as a basis for inferring processing differences between words and nonwords, is that word and nonword responses were not counterbalanced in this pilot study. A nonword response was the appropriate response for only one-fourth of the trials. Thus, subjects probably had a strong bias to respond word on any given trial. This does not undermine the important finding of this pilot -- subjects can perform this task with relative ease and speed.

When subjects were asked if they felt that the word/nonword judgment was difficult or if it interfered with their understanding of the sentences, the overwhelming response was no. An examination of subjects performance on the comprehension test confirmed their intuitions. The percentage of subjects who passed the comprehension test in this experiment (that is, made six or fewer errors) was exactly comparable to the percentage obtained for subjects who, in another experiment (Foss & Blank, in press) heard the same tapes, were given the same comprehension test, but, were given phoneme monitoring instructions.

It is noteworthy that in the Foss & Blank study phoneme monitoring reaction times to targets contained in real words were no faster than reaction times to targets contained in nonsense words. Foss and Blank point out that this finding calls into question the assumption that phoneme monitoring latencies reflect lexical processing of the target-bearing word. In contrast, the results of the present pilot study suggest that phoneme-triggered-lexical-decision can tap into lexical access at a point during sentence processing when phoneme monitoring cannot. That is, the task demands of phoneme-triggered-lexical-decision appear to ensure lexical access. A stronger test of this claim follows.

Main Experiment

On the basis of the preliminary findings of the pilot study, a more rigorous investigation of the usefulness of PTLD as a measure of lexical access during sentence processing was carried out. It was decided that an appropriate test of the sensitivity of this task to lexical access would be to manipulate semantic context. Numerous studies have found that the identification of a word in a list is facilitated by the prior occurrence in the list of semantically related words (e.g., Fischler, 1977; Meyer & Schvaneveldt, 1971). In addition, a few studies indicate that a similar effect can be observed in sentences. Blank and Foss (1978) have presented evidence suggesting that lexical access during sentence processing is facilitted by the prior occurrence of semantically related words within a sentence. This result was replicated by Foss, Cirilo & Blank (1979). Morton and Long (1976) have also argued that lexical access is faster when a relevant semantic context has occurred earlier within a sentence. In their study the semantic context typically could not be pinpointed to a specific word or two. Rather, it was the interpretation of the sentence fragment that seemed to affect lexical processing.

The main study reported here was designed to test the hypothesis that phoneme-triggered-lexical-decision response times reflect the time needed for lexical access as it occurs during sentence processing. If this is the case, response times should be sensitive to the facilitation of this process by the prior occurrence of a semantically related word. In order to test this hypothesis, the prior occurrence of a semantically related word was varied within a set of experimental sentences. In each sentence the verb was either related or unrelated to the noun that served as its direct object. Hence, an experimental sentence had two versions which differed from each other by the occurrence (or nonoccurrence) of a semantically related word. This is illustrated by the two example experimental sentences shown in Table 2. In each of the experimental sentences the direct object was the target word. Consequently, word/nonword judgments were triggered by the initial phoneme of the direct object. In the first example sentence, the target word is "car." If phoneme-triggered-lexical-decisions are sensitive to accessing the target word from the mental lexicon, then PTLD response times should reflect the facilitative effect of the semantically related word "drove" on the lexical processing of "car" relative to the effect of the unrelated word "headed."

Insert Table 2 about here

An important difference between this experiment and the pilot study is that the comparison of interest is not between WORD and NONWORD responses. Instead, responses to the experimental sentences are always WORD judgments and the important comparison is between WORD judgments for words preceded by a related verb and WORD judgments for words preceded by an unrelated verb. The function of the nonword judgments made on filler sentences is to make the task a sensible one for subjects to perform.

Table 2

Example sentences used in

P.T.L.D. follow-up study

	Trigger phoneme	Sentence
(1)	/k/	The student $\begin{Bmatrix}\text{headed}\\\text{drove}\end{Bmatrix}$ the expensive car back into a tree when the instructor wasn't looking.
(2)	/b/	The professor $\begin{Bmatrix}\text{finished}\\\text{published}\end{Bmatrix}$ his controversial book during the spring.

Method

Design and Materials: Twenty basic experimental sentences were constructed. Each sentence had two versions which defined the two experimental conditions: the verb was either related or unrelated to the direct object. In order that each basic sentence could occur in both conditions across the experiment, two material sets were constructed. Each set contained all 20 basic sentences; 10 of the sentences in both sets came from each of the two conditions. The experiment was therefore a 2 (verb context: related/unrelated) X 2 (material sets) factorial design with the former variable being within-subjects and the latter being between-subjects.

Thirteen of the related verb-noun pairs used in the experimental sentences were selected on the basis of relatedness ratings obtained from 127 undergraduate psychology students who did not participate in the main experiment. The subjects doing the rating were given a list of 100 simple sentences, all of the type: (Det) N V Det (Adj) N. In each sentence either the verb or the adjective was underlined along with the direct object. The subjects' task was to judge how related the two underlined words were to each other. More specifically, subjects were asked to judge how often the first underlined word made them "think of" the second underlined word. Each subject made 50 verb-noun pair judgments and 50 adjective-noun pair judgments. Subjects used a five-point rating scale to indicate their judgments, where 1 represented 0-20% of the time and 5 represented 80-100% of the time. Only those verb-noun pairs for which the degree of relatedness of the verb to noun was judged to be between 70-100% by a minimum of 75% of the subjects were used in the main experiment. The remaining seven related verb-noun pairs used in the experimental sentences were those which consistently elicited a particular noun as the response when presented to several colleagues for judgment. On the other hand, there was no agreement among noun responses when the unrelated verbs were presented. The frequency of the related and unrelated verbs were matched according to Kucera and Francis (1967) estimates.

The initial structure of each experimental sentence was NP V Det Adj N ... In these sentences the direct object noun was always the target word; hence, word/nonword judgments were triggered by the initial phoneme of the direct object. Five stop consonants were used as targets among the experimental sentences with the following frequencies of occurrence: /b/-6; /p/-4; /d/-3; /t/-2; /k/-5. Twenty filler sentences were constructed which had nonwords as the target "words." That is, the target phoneme occurred on a nonword. Six stop consonants were used as targets among these "nonword" fillers with the following frequencies of occurrence: /b/-2; /p/-4; /d/-2; /t/-4; /g/-4; /k/-4. Ten additional fillers were constructed which did not contain a target word beginning with the specified target phoneme. Of these "non-target" fillers, five contained a nonword. A final group of 10 fillers were constructed which had real words as the target word. Of these "real word" fillers, one half contained a nonword. Table 3 shows examples of the various types of filler sentences used in the experiment. The filler sentences were identical for both material sets. The total 60 sentences were randomized, with each basic experimental sentence occurring in the same position for the two sets of experimental materials.

Insert Table 3 about here

A female speaker recorded the two material sets on one track of an audio tape. A pulse, inaudible to subjects, was placed on another track of the tape at the beginning of the initial phoneme of each target "word." The pulse started a timer which stopped when subjects pressed a button. Timing and data collection were controlled by a PDP 8/I computer.

Subjects: The subjects were 37 undergraduate psychology students at the University of Texas at Austin who participated in the experiment in partial fulfillment of a course requirement. Twenty subjects were assigned to one material set; 17 were assigned to the other.

Table 3

Filler sentences used in

follow-up study

# of fillers	Trigger phoneme	Sentence	Correct response
20	/b/	The girls finally decided to throw out the brebben clock since it was obviously not worth repairing.	NONWORD
5 (2 with nonword occurring after the target; 3 with nonword occurring before)	/t/	After entering the tournament the golfer broke his dezzer and withdrew from the competition.	WORD
5	/g/	Last night the girl next door took the huge dog out for a walk.	WORD
5	/p/*	The man in the restaurant ordered a large frezmon for dinner	NONE
5	/d/*	Yesterday afternoon a U.F.O. was sighted over New York City.	NONE

*target phoneme did not occur in the sentence

Procedure: Subjects were tested in groups of one to six, with the experimenter and subject occupying adjoining rooms. Each subject was seated in a booth out of direct sight of the others.

Instructions describing the subjects' task were recorded at the beginning of each experimental tape. The instructions and the test sentences were presented binaurally over headphones. Subjects were told to lightly rest the index finger of each hand on the two response buttons in front of them. One button was labelled real word; the other was labelled nonsense word.

Subjects were informed that they were going to be presented with a list of sentences and that some sentences would contain a nonsense word and some sentences would contain only real English words. Their instructions were to press one of the buttons as quickly as possible when they heard a "word" in the sentence that began with a particular target sound (i.e., "buh as in Bob"). More specifically, subjects were told to press the real word button if the word beginning with the target sound was a real word, and to press the nonsense word button if the word beginning with the target sound was a nonsense word. A trial consisted of the word "ready," specification of the target phoneme, and then presentation of a test sentence.

Subjects were also told that the target sound for those sentences that contained a nonsense word would not always occur on the nonsense word and that they should not allow the presense of a nonsense word to interfere with their task of pressing the appropriate button when they heard a word beginning with the target sound. In addition, subjects were told that some sentences would not contain a "word" beginning with the target sound specified for that sentence. In these cases neither button was to be pressed. Subjects were given four practice sentences: one contained a real word target in a sentence with only real English words; one was a sentence with a nonsense word target; one contained a nonsense word but the target was a real word; and one was a sentence with all real English words but without a target-bearing word.

Subjects were forwarned in the instructions that sometimes, immediately after they heard a sentence a yes/no question would be asked pertaining to that sentence. In addition, they were told that they would be able to answer correctly only if they had understood the sentence. The importance of paying close attention to the sentences was emphasized by the fact tht listeners did not know beforehand which sentences in the list would be followed by a question. Subjects indicated their answers to the comprehension questions by circling either YES or NO on the appropriate line of a printed answer sheet. A total of 16 questions were presented. Table 4 presents two sentences which subjects were tested on along with the corresponding questions and correct answers. The same comprehension test was administered to all subjects. A fifth practice sentence was given which was followed by a comprehension question. After the experimenter answered questions clarifying any uncertainties regarding the instructions, the list of experimental and filler sentences was presented.

Insert Table 4 about here

Results

The mean RTs for each subject in the two experimental conditions were computed. The results for both conditions are shown in Table 5. These reaction times have been truncated in the following way. A mean and standard deviation was computed for each subject and for each item in the experiment. If any individual RT was more than 2.5 standard deviations from both the mean for the subject and the mean for the item, it was omitted and replaced by a procedure suggested by Winer (1971). This procedure resulted in replacing about two percent of the data. Missing data points which resulted from failures to respond (about five percent) were also filled in according to Winer's procedure.

Insert Table 5 about here

Table 4

Sample comprehension questions

(1) Sentence: All the major networks covered the lunar bouving yesterday afternoon.

Question: Was yesterdays event of national interest?

Answer: YES

(2) Sentence: After hearing all the testimony, the jury found the defendant not guilty.

Question: Should the defendant be imprisoned?

Answer: NO

Table 5

Mean reaction times (msecs)

related/unrelated verb experiment*

Related Verb	Unrelated Verb
736	783

*Note. Responses to the experimental sentences were all WORD judgments.

An unequal N analysis of variance by subjects showed a significant main effect for the verb context, $F_1(1,35) = 9.23$, $p < .004$. The analysis of variance by items also showed a significant verb context effect, $F_2(1,19) = 5.07$, $p < .03$. No other main effects or interactions were significant. Error rates for the two conditions were: related verb -- .04; unrelated verb -- .06. The results of a t-test for dependent samples indicated that these error rates did not differ significantly from each other, ($t_{36} = 1.96$, $p > .05$).

The mean number of errors per subject on the comprehension test was .51. The overall error rate for performance on the comprehension test was less than 3 percent, $p < .001$ of obtaining this level of performance by chance. The small number of errors on the comprehension test suggests that phoneme-triggered-lexical-decisions do not interfere with comprehension per se.

Figure 1 shows mean reaction times for the 20 experimental sentences as a function of trial number. There appears to be no systematic practice effects. Although there is an initial increase in response times, latencies do not remain at about 850 msecs. Instead, they seem to fluctuate randomly between 670 and 900 msecs. The fact that reaction times do not demonstrate a trend of any kind over trials suggests that subjects are not adopting unusual processing strategies as they become more familiar with the task.

Insert Figure 1 about here

Discussion

The finding that word judgments for words were faster when they were preceded by a semantically related verb than by an unrelated verb supports the hypothesis that phoneme-triggered-lexical-decision response times reflect on-line lexical processing during sentence comprehension. The basis for this conclusion is, perhaps, more readily apparent if one considers the results of this study in conjunction with Morton's (1969) proposed model of word recognition.

According to Morton's model, each entry in the mental lexicon (or logogen) has associated with it a threshold value which defines the amount of information which must be received for the logogen to be activated or accessed. Both sensory and semantic information are accepted as input to the logogen system. These two types of information combine to activate logogens in such a way that there is a trade-off relationship between them. More specifically, when a logogen receives semantic input from prior relevant context there is a decrease in the amount of sensory information that is needed to bring it above its threshold.

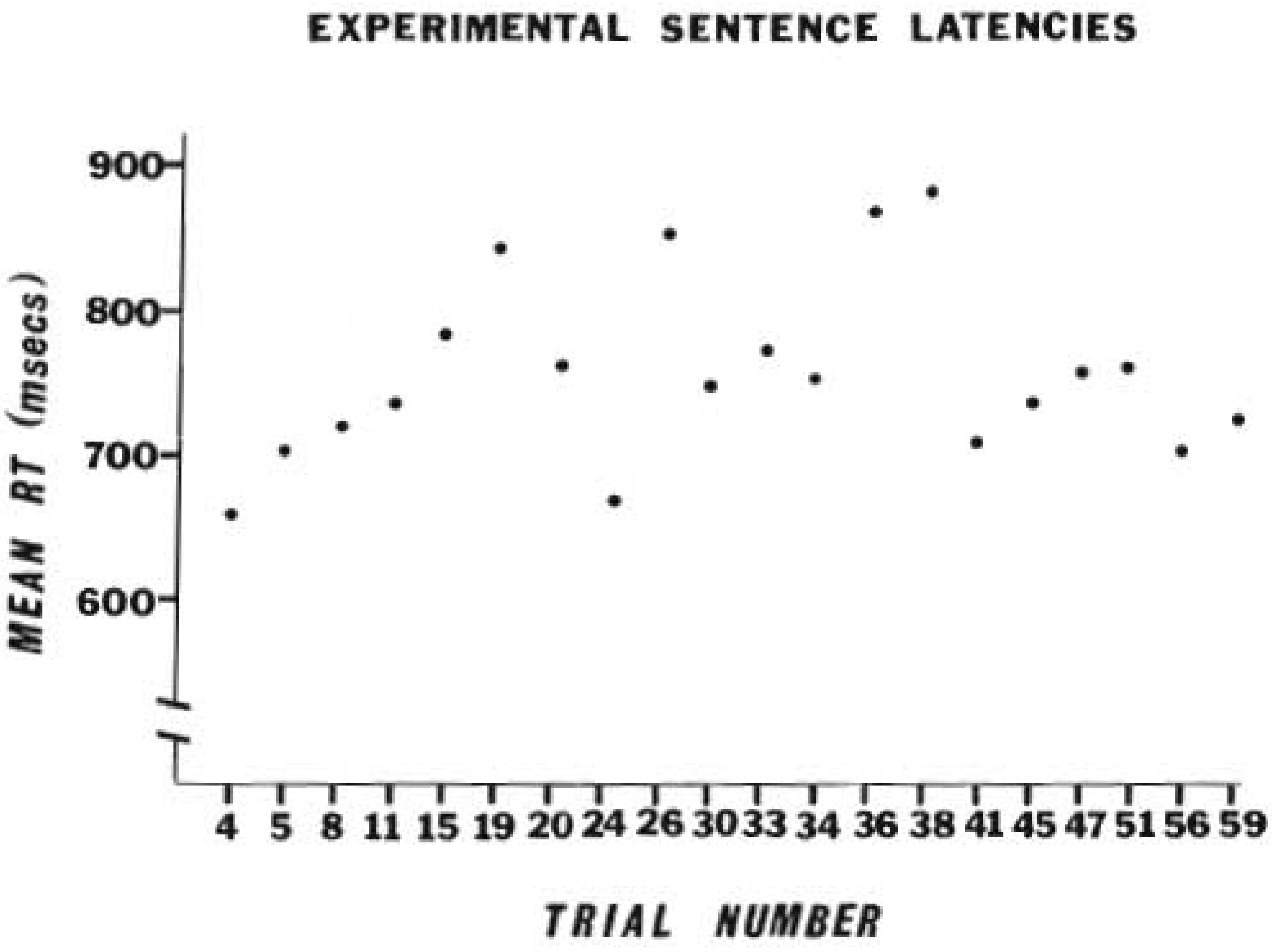

EXPERIMENTAL SENTENCE LATENCIES
MEAN RT (msecs)
900
800
700
600
4
5
8
11
15
19
20
24
26
30
33
34
36
38
41
45
47
51
56
59
TRIAL NUMBER

figure 1

The results of this experiment are easily interpretable in accordance with Morton's logogen model. The spoken waveform corresponding to a target word required less sensory processing when it was preceded by a semantically related verb as compared to when it was preceded by an unrelated verb. Presumably, this is because the logogen corresponding to the target word had been partially activated by the contextual information provided by the previously accessed semantically related verb. Spreading semantic excitation among related lexical concepts in the mental lexicon is one way for semantically related words to partially activate logogens (see, e.g. Collins & Loftus, 1975). The finding of semantic facilitation in this study is consistent with two earlier studies showing context effects on word recognition during sentence processing (e.g., Blank & Foss, 1978; Morton & Long, 1977). Since phoneme-triggered-lexical-decision response times are sensitive to the facilitation effects of semantic relatedness on lexical processing, it seems reasonable to conclude that this measure indeed reflects lexical access as it occurs during sentence comprehension.

If investigators are interested in studying the effects of various factors on lexical access during sentence processing, it is imperative that they have an understanding of the sensitivities and limitations of their measurement techniques. Unlike phoneme monitoring, the task demands of phoneme-triggered-lexical-decision guarantees that a target-bearing word will be lexically processed prior to execution of the response. Phoneme monitoring latencies may provide inaccurate measures of lexical access since it cannot ensure processing at this level -- responses are sometimes initiated before lexical access has occurred.

Similar criticisms may also be relevant in evaluating the merits of the "word monitoring" task (e.g., Marslen-Wilson & Tyler, 1975) as an on-line measure of lexical access during sentence processing. In particular, it may be that the task specifications of word monitoring introduce contextual factors which confound the time needed for lexical access. We know that the prior occurrence of a semantically related word affects speed of lexical access. Accordingly, it is more than likely that the prior occurrence of the exact word (as the target specification) in word monitoring will affect the ultimate processing of the target word. If this is the case, it is important that such effects be taken into account in interpreting word monitoring response times. A systematic investigation of phoneme-triggered-lexical-decision, phoneme monitoring, and word monitoring as measures of lexical access during sentence processing is currently underway.

One final point should be made before summarizing. The phoneme-triggered-lexical-decision task lends itself to slight modifications which may allow component processes other than lexical access to be measured during sentence comprehension. By simply changing the nature of the judgment triggered by the target phoneme one may be able to tap directly into, say, inferential processes. The "on-line" construction of inferences could be investigated by having subjects make decisions that would require the making of an inference. Essentially this task could be used to trigger virtually any type of judgment during sentence processing. Moreover, by changing the location of the target within a sentence one could study the time course of these sub-processes. However, when using this or any other type of judgment task the following two factors should always be considered: first, the relationship between the kind of judgment being made and the comprehension process; second, the inherent difficulty of the judgment being made. These two factors determine how the task demands might interact with sentence processing mechanisms, and this, in turn, affects the possible conclusions that can be drawn from the observed results.

To summarize, a new on-line measure of lexical access during sentence processing has been proposed. The results of the present PTLD study provides additional support for the facilitative effects of semantic relatedness on word recognition in fluent speech.

Reference Note

1. Marslen-Wilson, W.D. & Tyler, L.K. The temporal structure of spoken language understanding. Manuscript submitted for publication.

References

Aaronson, D. Performance theories of sentence coding: Some qualitative observations. *Journal of Experimental Psychology: Human Perception and Performance*, 1976, *2*, 42-55.

Ammon, P.R., Ostrowski, B., & Alward, K. Effects of task on the perceptual organization of sentences. *Perception & Psychophysics*, 1971, *10*, 361-363.

Blank, M. A. & Foss, D. J. Semantic facilitation and lexical access during sentence processing. *Memory & Cognition*, 1978, *6*, 644-652.

Carey, P.W. Verbal retention after shadowing and after listening. *Perception & Psychophysics*, 1971, *9*, 79-83.

Cole, R.A. Listening for mispronunciations: A measure of what we hear during speech. *Perception & Psychophysics*, 1973, *3*, 153-156.

Collins, A.M. & Loftus, E.F. A spreading-activation theory of semantic processing. *Psychological Review*, 1975, *82*, 407-428.

Fischler, I. Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, 1977, *5*, 335-339.

Forster, K.I. & Chambers, S.M. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 627-635.

Foss, D.J. Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, 1969, *8*, 457-462.

Foss, D.J. Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, 1970, *9*, 699-706.

Foss, D.J. & Blank, M.A. Identifying the speech codes. *Cognitive Psychology*, in press.

Foss, D.J., Cirilo, R.K. & Blank, M.A. Semantic facilitation and lexical access during sentence processing: An investigation of individual differences. *Memory & Cognition*, 1979, *7*, 346-353.

Foss, D.J. & Jenkins, C.M. Some effects of context on the comprehension of ambiguous sentences. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 577-589.

Foss, D.J. & Lynch, R.H., Jr. Decision processes during sentence comprehension: Effects of surface structure on decision times. *Perception & Psychophysics*, 1969, *5*, 145-148.

Foss, D.J. & Swinney, D.A. On the psychological reality of the phoneme: Perception, identification, and consciousness. *Journal of Verbal Learning and Verbal Behavior*, 1973, *12*, 246-257.

Kucera, H. & Francis, W.N. *Computational analysis of present day American English*. Providence, Rhode Island: Brown University Press, 1967.

Marslen-Wilson, W.D. Linguistic structure and speech shadowing at very short latencies. *Nature*, 1973, *244*, 522-523.

Marslen-Wilson, W.D. Sentence perception as an interactive parallel process. *Science*, 1975, *189*, 226-227.

Marslen-Wilson, W.D. & Welsh, A. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 1978, *10*, 29-63.

Meyer, D.E. & Schvaneveldt, R.W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 1971, *90*, 227-234.

Meyer, D.E., Schvaneveldt, R.W. & Ruddy, M.G. Loci of contextual effects in visual word recognition. In P.M.A. Rabbit & S. Dornic (Eds.), *Attention and Performance V*. London: Academic Press, 1975.

Morton, J. Interaction of information in word recognition. *Psychological Review*, 1969, *76*, 165-178.

Morton, J. & Long, J. Effect of word transitional probability on phoneme identification. *Journal of Verbal Learning and Verbal Behavior*, 1976, *15*, 43-52.

Newman, J.E. & Dell, G.S. The phonological nature of phoneme monitoring: A critique of some ambiguity studies. *Journal of Verbal Learning and Verbal Behavior*, 1978, *17*, 359-374.

Scarborough, D.L., Cortese, C., & Scarborough, H.S. Frequency and repetition effects in lexical memory. *Journal of Experimental Psychology: Human Perception and Performance*, 1977, *3*, 1-17.

Studdert-Kennedy, M. The perception of speech. In T.A. Sebeok (Ed.), *Current trends in linguistics, (Vol. XII)*, The Hague: Mouton, 1974, 2349-2385.

Winer, B.J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1962.

Footnotes

*This article is based on a paper originally presented at the Southwestern Psychological Association, New Orleans, 1978. I thank my friends and colleagues Don Foss, Wendy Idson, Pam Holley-Wilcox, and Dave Pisoni for their time in reading this manuscript and their many helpful suggestions.

III. INSTRUMENTATION AND SOFTWARE DEVELOPMENT

Speech Perception Laboratory: Current Computer Resources*

Jerry C. Forshee

Indiana University

Introduction

The current laboratory computer configuration has changed quite significantly since our last report on its status, see Forshee (1976). The original computer system was a DEC PDP 11E05 packaged system consisting of the PDP 11/05 processor, 16K core memory, a 2.5 Mbyte disk system, a 30 cps hardcopy terminal, and a dual cassette tape system. This computer system was delivered in the Fall of 1974 and was purchased to support our early work under NIMH research grant MH-24027. The primary instrumentation objective of this initial system was to support the presentation to human observers of very short (300 to 500 msec.) speech signals in real-time and the collection of their subsequent responses to these stimuli. To achieve this goal we developed two custom interfaces to support the required analog signal processing and the management of subject I/O (i.e., cue and feedback lights on the output and various types of response inputs). These interfaces have been described in some detail in earlier progress reports (see Forshee 1975, 1976) and the interested reader is urged to consult these for further information.

Our success with this initial system was sufficient to allow support of a parallel although separate research project in 1975 using this computer system as a basis of our instrumentation needs. Our request subsequently led to funding from NINCDS under research grant NS-12179. The acquisition of this second grant allowed us to expand the original computer hardware system and along with it our capacity for conducting on-line perceptual experiments (see Forshee 1976). At this time, we were also able to give some emphasis in our laboratory toward developing an initial program library for supporting synthetic speech generation and some initial steps toward digital signal processing techniques and rudimentary computer graphics. The equipment added to the system during 1975-1976 under this research project included: a CRT terminal, a second 2.5 Mbyte disk, a VR-14 point plot scope, an AR11 analog and clock interface, an additional 12K of memory, an expander box, a high speed magnetic tape system and an extended arithmetic element.

In the past several years the PDP 11/05 computer system has not seen any major changes like these initial two stages of its development, except for the addition of an occasional needed peripheral device. These gradual additions have included the acquisition of a digitizing tablet from research funds provided by Indiana University in 1977, a VT-11 graphics processor and display system in 1977 and a Tektronix scope and hardcopy unit which were purchased with funds from grant NS-12179 in 1979.

These later changes have reflected our desire to support somewhat more sophisticated digital signal processing requirements, including: graphic waveform displays, digital waveform editing, spectral analysis of speech via LPC and the generation of synthetic speech signals via digital synthesis techniques.

As our interest in these capabilities grew, and as our library of application programs to support it expanded substantially, it became obvious that the PDP 11/05 was going to be inadequate to handle all of our computational and experimental needs in the future. When it became clear that the laboratory needed more processing power, we were able to obtain additional funds under a renewal application to NIMH and we purchased a separate PDP 11/34 computer system which was delivered in the Fall of 1978. This system consisted of: an 11/34 processor, a 32K word MOS memory system, a 7.5 Mbyte disk system and a 30 cps hardcopy terminal. Since the arrival of this new computer system several add-on options have been made to its initial configuration. During this same period the laboratory has also undergone a major physical reorganization to meet the space needs of our new computer facilities. Our desire was to configure the laboratory with the PDP 11/05 performing mostly "on-line" perceptual experiments and the PDP 11/34 doing the computationally bound signal processing/analysis and synthesis tasks (see Kewley-Port, 1978; 1979). Wherever possible, however, we have attempted to overlap functions of both computers so as to achieve the greatest possible laboratory throughput. In the following sections of this report we will describe in some detail the design goals, the current configuration and typical functions supported by each of the computer systems currently in the Speech Perception Laboratory. This report is not an exhaustive description of these systems and is simply intended to summarize the current capabilities of the laboratory at the present time for carrying out a fairly wide range of activities associated with speech processing tasks including analysis, synthesis and perceptual experimentation.

The PDP 11/05 Computer System

System Design Goals.

Our primary functional objective for the PDP 11/05 was to have a streamlined computer system capable of conducting the bulk of our on-line perceptual experiments. Our definition of conducting perceptual experiments includes not only the real-time presentation of test signals to observers and the collection of their responses but also the development of all required programs and the data recovery tasks, i.e., data summarization and analysis, as well. With this objective in mind we fully realized that particularly demanding experiments could not be conducted on the PDP 11/05 and would necessitate access to the PDP 11/34 system. We do not view this as a limitation in the laboratory at the present time but only a most parsimonious division of labor. To achieve this goal it was necessary to evaluate the individual elements and components from the previous configuration accessing their function as being best suited towards conducting perceptual experiments or some aspect of digital speech processing. When an equipment item could be placed in this latter category it was removed from the PDP 11/05 system and reinstalled on the PDP 11/34 system. The resulting configuration from the application of these criteria is summarized graphically in Figure 1.

Insert Figure 1 about here

This configuration can be roughly divided into five functional divisions: (1) basic system resources, (2) analog output system, (3) subject I/O and monitoring, (4) video display system, and (5) word processing. These divisions will be summarized below.

Basic System Resources.

After the previous configuration was paired down to those peripherals that best suited our needs for conducting on-line perceptual experiments, the system's central computational resources consisted of: the KD11-B processor which includes an integral KW11-L line frequency clock; 28K words of central memory

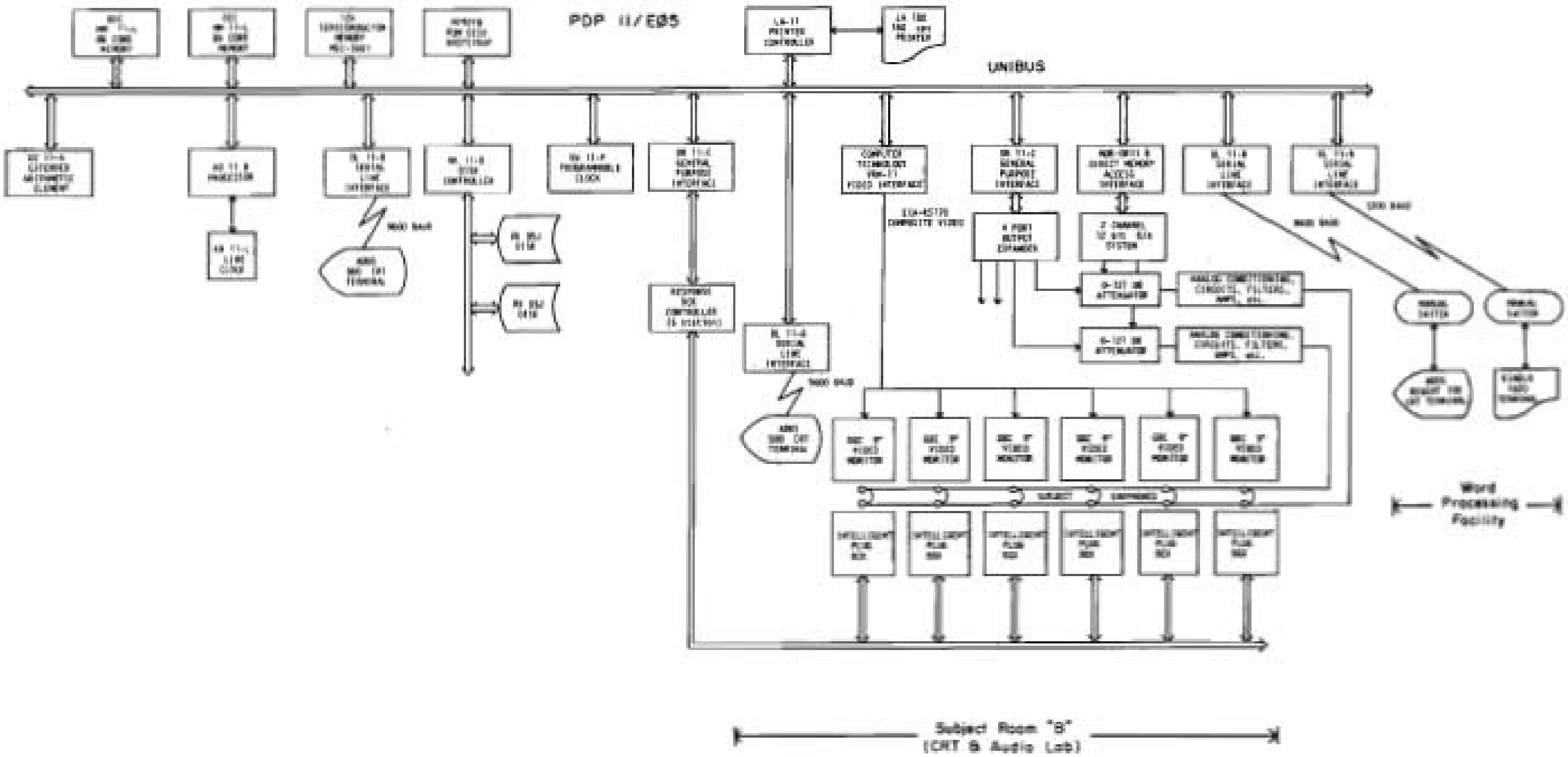

PDP 11/EØ5
UNIBUS
LA-11 PRINTER CONTROLLER
DR 11-C GENERAL PURPOSE INTERFACE
COMPUTER TECHNOLOGY VDA-11 VIDEO INTERFACE
EIA-RS170 COMPOSITE VIDEO
4 PORT OUTPUT EXPANDER
2 CHANNEL 12 BIT D/A SYSTEM
0-127 DB ATTENUATOR
ANALOG CONDITIONING CIRCUITS, FILTERS, AMPS, etc.
SERIAL LINE INTERFACE
9600 BAUD
1200 BAUD
MANUAL SWITCH
VIDEO MONITOR
INTELLIGENT PLUG BOX
SUBJECT
Word Processing Facility
Subject Room "B" (CRT & Audio Lab)

Figure 1. Current configuration of the PDP 11/05 computer system in the Speech Perception Laboratory.

including 16K of DEC MM11-L core and 12K of Monolithic Systems MOS memory; and the KE11-A, an extended arithmetic element. The central peripheral resources of the system now consist of: a high speed (9600 baud) ADDS 980 CRT terminal, the 5 Mbyte hard disk system consisting of the DEC RK11-D controller and two RK05-J drives featuring removable media; the disk system can also logically be thought to include the M792YB ROM bootstrap loader, hardcopy is provided by the very cost effective LA180 character printer which operates at 180 cps, and finally the various software timer functions which are driven by the hardware KW11-P programmable real-time clock.

Analog Output System.

In our first effort to support input and output of speech signals under computer control, we chose a hardware scheme that was highly integrated using a design that was not very flexible (see Forshee, 1975). As is always the case, experience was a very profound teacher and we found this to be an unsatisfactory hardware organization for implementing a wide range of activities. The original system was generally reliable and it allowed for several years of high volume output and perceptual experimentation on speech. But our main problem was that the hardware system was closed, fixed and not readily adaptable to change as our experiment to experiment instrumentation demands evolved over the last few years. As such dynamic growth is the logic and analog power systems are therefore kept totally isolated. The only two signals leaving the computer are the two D/A analog outputs which use isolated grounds and differential line drivers for maximum isolation and noise immunity. These two analog signals are connected to a separate equipment rack containing all the analog signal conditioning circuits. The ultimate design criteria for the analog rack was one of modularity and signal standardization. Each module has a standard input/output impedence and voltage protocol. This design scheme allows for modules to be patched in or out of the signal path quickly to satisfy the need of a particular experiment or related task. This modular standardization also allows for quick repairs, usually by exchanging the defective module with a spare one on hand. Such standardization also reduces the effort of incorporating new devices in the system; all that is needed is to convert the new device to our standard, if required, before adding it to the analog equipment rack. Another advantage is the great flexibility it permits; existing signal conditioning functions may be removed and new ones added in the chain with no effort required to break the chain or rearrange the existing modules. The standard units currently employed in this rack include: (1) the differential line receiver module which receives the signals from the D/A interface in the computer and establishes the electrical protocol for the analog signals; (2) low pass filters, both 5K and 10K are available; (3) a programmable attenuator, for each analog channel which has 9 stages with attenuation step sizes of 1/4 dB, 1/2 dB, or 1 dB, our default configuration is 1 dB steps with a range of 0 to -127 dB; (4) auxiliary output panel for driving experimenter earphones or speakers and other output devices such as tape recorders or the sound spectrograph; (5) the final link in the analog interface chain is the power amplifier which drives a series of earphones placed in each experimental subject booth.

The digital outputs required to operate the attenuators come from a simple yet extremely useful device we designed and constructed to provide an extended number of output bits while imposing a minimal load on the Unibus. This interface, the DR11-C Output Expander, produces four words of digital output with a single load presented to the Unibus by using a standard DR11-C module. The one word output of the DR11-C is multiplexed to four 16 bit holding registers controlled by two bits of the DR11-C control and status register. One of these expanded output words is used for controlling each of the programmable attenuators.

Subject I/O and Monitoring.

As we learned a great deal from our first attempt with analog interfacing to the computer, we also learned from our first attempt at interfacing the human observer to the computer. The subject Response Box Controller (RBC) described in our earlier report (see Forshee, 1975) used a binary coded scheme for both address and data busses. We found that this system contained several inherent problems making it less desirable than an unencoded parallel data buss structure. As a consequence, we constructed a new version of the RBC for the present system. The RBC is basically a data multiplexer which is connected to the computer via a DR11-C interface. Multiplexing is provided to allow an output buss of eight data bits to each of six subject stations, each one being independently latched in the RBC. Likewise, on the input, the RBC acts as a holding register for eight bits of input data for each of these six subject stations. Digital logic is provided in the RBC to handle the stacking of input data in the case of more than one subject station responding simultaneously. When multiple responses occur, the data is held in the RBC and an output interrupt to the computer is generated each time an input is received until all pending responses have been read by the computer. In this second attempt at interfacing the human observer, a more versatile and modular approach was taken to the organization of electrical signals available at the subject station. A new device called the Intelligent Plug Box (IPB) was also designed which greatly facilitates the frequent changes between particular response manipulanda demanded by different experimental paradigms in use in the laboratory.

The IPB provides the interface between the RBC and the several types of response boxes used at each subject station. The

IPB provides four different jacks that connect subject I/O devices using two different electrical protocols. The first connector jack provides points to connect the eight input and eight output lines using the "event" protocol. This protocol supports switches, pushbuttons, small relays, incandescent lamps and similar devices. The second output connector repeats the output lines so that input and output functions can be built in separate enclosures and then mixed and matched to suit a particular experiment. For "event" devices no additional logic is required beyond that contained in the IPB itself. Thus, it is quite economical to have multiple sets of subject I/O devices as only the enclosure, buttons, and cue lamps need be duplicated. This organization also makes it quite possible for a laboratory user to construct a unique subject I/O device for his particular needs with a minimal amount of technical knowledge and skill and with minimal time requirements.

A second signal protocol is used on the third and fourth connectors on the IPB. The standard TTL logic family protocol is used with both the eight input and eight output lines appearing on the third connector while only the eight output lines are repeated on the fourth, for the same reason of combining I/O boxes as in the "event" protocol. This second mode, the "logic" protocol, is used to directly connect keyboards, numeric keypads, LEDs, seven segment displays or custom devices that use conventional TTL logic signal protocol.

Subject monitoring during experiments is carried out by a high speed CRT terminal located in the subject room. The experimenter usually remains in the subject room during an experimental session and uses this terminal to start, pause and run the actual on-line experiment. When appropriate, the experimenter can enter new parameter values as required in conducting the experimental session. This terminal is used extensively by the particular program conducting the experiment to provide data to the experimenter regarding the progress of each individual subject participating in the experiment. This information can be used to eliminate entirely, or simply prompt a low performing subject, to motivate him to raise his responding level to the standards of the current experiment.

Video Display System.

Two requirements have led us to install in each subject booth a 9" video monitor (GBC Model MV10-A) for the presentation of alphanumeric text to subjects as part of the experimental process. This system can be used to present verbal instructions, feedback and general prompting to subjects on an individual basis. This system is also being used in some pilot studies in which a stimulus is presented both auditorily and in some visual representation simultaneously. The initial configuration has all of the subject stations driven by a single video generator device produced by Computer Technology (Model VRU-11). This device is capable of generating a video display of 25 lines of alphanumeric text each consisting of up to 80 characters. This system has been designed with considerable flexibility in mind, so that additional video generators can be added later for experiments requiring a larger number of characters in the display, and larger display capacity required by each subject having his own independent display. This system will be able to provide up to a maximum of one full 25 line by 80 character display for each of six subject stations running independently in this laboratory.

Word Processing.

In the past year or so we have found that using the computer system for word processing in off hours and at other times when it is not being used in research oriented tasks is a tremendous labor saving device. Using the computer for word processing is very effective way for researchers in the laboratory to type and edit their own scientific reports and scholarly papers. These same facilities are employed by the technical staff of the laboratory in preparing internal documentation on the use of various new and refined hardware and software facilities available in the laboratory. Toward this end we have established a corner of the laboratory as the word processing area. In this work space, we have placed an upper/lower case CRT terminal and a high print quality Diablo 1620 terminal. These two terminals are wired to a junction box that contains the necessary logic

The PDP 11/34 Computer System

System Design Goals.

Given the financial limitations of our research activities our goal for the PDP 11/34 system was to configure the fastest computational machine possible for performing a variety of digital signal processing tasks. As a secondary goal, we decided that this system should also be capable of conducting perceptual experiments where the demand on the computer during stimulus presentation was very extensive (i.e., in adaptative testing situations). Thus, we sought to establish a computer system that was upward program compatible with the PDP 11/05 system, so that all of our previous program development could be directly transferred to the new system; and to add all those options to the computer which provide increased computational throughput. Although considerable effort has been directed toward the design of this system, it is obvious that the current configuration is not completely permanent; as new research needs are encountered modifications will need to be made to meet these needs. What we are certain of, however, is that the PDP 11/34 system is an excellent choice around which to build a flexible laboratory of this type for it provides a very adequate set of facilities in its basic configuration and allows for more add on capability than any small-sized research laboratory would ever have need to add. The current configuration of the PDP 11/34 computer system is displayed in Figure 2, and will be described below under the following major headings: (1) basic system resources, (2) analog input/output system, (3) digital signal processing (4) subject I/O monitoring, and (5) word processing.

Insert Figure 2 about here

Basic System Resources.

The basic processor in this system has been complemented by two add-on options which combine to make the PDP 11/34 a very fast computational machine. These include the KD11-EA processor itself, the FP11-A floating point processor and the KK11-A cache memory. This later device acts as a high speed buffer between memory and the processor, providing a 40% to 60% increase in throughput depending on the organization of the program being executed. The total memory on the system is now 96K words which consists of the original DEC 32K word MS11-EJ and a Monolithic Systems 64K word MOS 3603 system. The processor is equipped with the optional Programmers Octal Keypad Console, the KY11-LB subsystem. The system programming console is an ADDS 580 CRT terminal, operating at 9600 baud, and connected to the Unibus by a DEC DL11-W serial line interface which also provides the hardware line frequency clock function. The system storage device

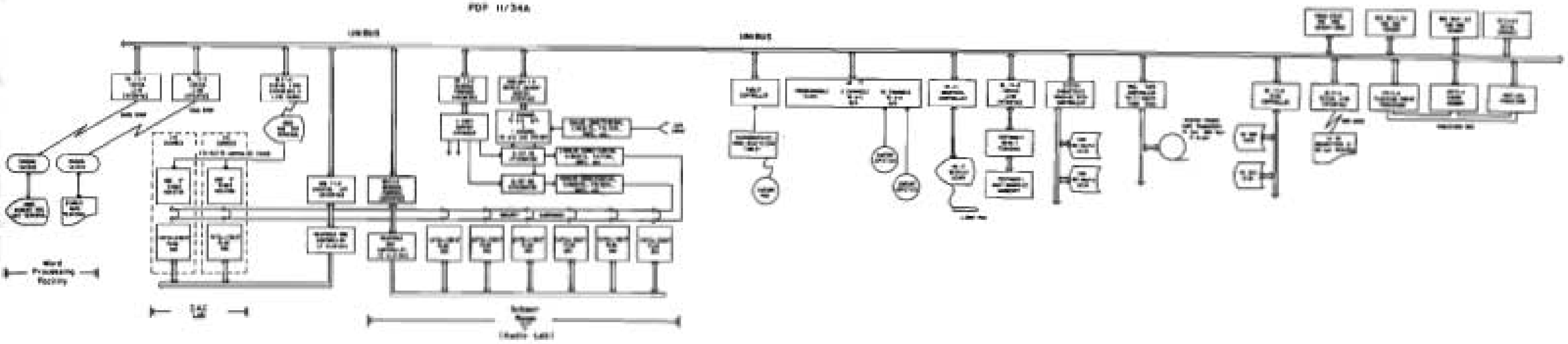


Figure 2. Current configuration of the PDP 11/34 computer system in the Speech Perception Laboratory.

is an RK05 disk system consisting of the DEC RK11-D disk controller, an RK05-J removable media 2.5 Mbyte drive and a fixed media 5.0 Mbyte RK05-F drive. The system hardcopy device is a DEC LA-36 DECwriter which we are using for output only. This has been modified to print at a continuous 60 cps rather than the normal 30 cps. The programmable real-time clock on this system is one of the functional units of a DEC AR-11 interface. This clock is used to drive the various software timers used throughout the system.

Analog Input/Output System.

The analog subsystem is a direct functional superset of the analog output system used on the PDP 11/05 summarized earlier. This system is exactly equivalent on the D/A output side and differs only in that a single 12-bit channel of analog input (A/D) is provided for digitizing external analog signals. This A/D input channel has several modules of analog signal conditioning circuits used to perform filtering, switching, amplifying and attenuation functions which are all built with the same modularity and signal standardization goals that we described above for the D/A system on the PDP 11/05 computer.

This system also uses a DR11-C output expander interface to provide an extended number of digital output points. In addition to providing the bits necessary to control the programmable attenuators in each output channel, one word of this device is used to drive our OVE IIId Speech Synthesizer (see Sawusch, 1975). The OVE IIId is a hardware speech synthesizer that produces a single analog output in real-time in response to parallel digital input. This device was previously attached to the PDP 11/05 system (see Forshee, 1975). The output of the OVE can be made available at the analog equipment rack and therefore can be routed to any transducer destination as can the D/A analog signal sources.

Digital Signal Processing.

Digital Signal Processing is a somewhat generic title under which we have lumped all of our digital (program managed) processing activities including both speech and non-speech signals. This includes generation of synthetic stimuli using digital and algorithmic techniques, editing and display functions of signals in digital form, and spectral, LPC and temporal analysis of speech stimuli. To support these tasks the lab developed a number of very sophisticated application programs (see Carrell & Kewley-Port, 1978; Kewley-Port, 1978 and Kewley-Port, 1979).

The hardware support for our digital signal processing requirements consists of several rather specialized items of equipment. Programs that require interactive graphic displays use the DEC VT-11 graphics controller which produces graphic output on the VR-17 display scope. This system is also equipped with a light pen which can be used for interacting with the graphic displays or with the program generating the display. Using analog inputs on the AR-11 interface module, we have added two multiturn pots which are used as cursors with the VT-11 graphic display. Once the user has interacted with an application program and produced a graphic plot worth saving, the Tektronix equipment can be called into play. Using a software routine, a user can copy the display vectors directly from the VT-11 display system data base main memory to the Tektronix 4010 terminal. Once the display has been copied to the 4010 hardcopies can be produced easily on the Tektronix 4631 hardcopy unit. Data in two dimensional graphical representation, such as spectrograms produced by our Voice Print Sound Spectrograph, can be input to the PDP 11/34 computer system using a digitizing tablet. For this function, we have interfaced a Summagraphics 2000 series digitizing tablet and integral cursor pad to the computer.

Currently under development is a continuous digitizing system where long passages of speech can be digitized as one long continuous stream of digital information. In this mode, the digital signals can be directly buffered onto high-speed digital magnetic tape. Supporting this function is a Pertec T9000 series 9 track, 800 bpi tape transport which operates at 75 ips. This peripheral is connected to the Unibus by a Tano series 5100 magnetic tape controller which is compatible with the DEC TM11 series mag tape system. The mag tape system also sees considerable use as an off-line backup medium from disk for programs and experimental data.

We plan to add a very large disk system to the battery of peripheral devices in the lab. This system will be connected to the Unibus via a System Industries 9400 series controller which allows for up to eight 80 Mbyte disk drives to be added to the system. Our principle use for this large disk system will be storage, random access and retrieval of very large samples of speech such as sentences and paragraphs as well as synthesis by rule of connected fluent speech via a text-to-speech system.

Subject I/O and Real-Time Capabilities.

The real-time capabilities of interacting with human observers on the PDP 11/34 are quite similar to the facilities described earlier for the PDP 11/05. On this system we have an identical RBC system controlling our Audio Laboratory. In this subject room, we have six subject stations each equipped with two-channel earphones and an IPB, which allows us to run any one of our standard experimental paradigms. A smaller two-subject RBC system is used to interface IPB's in each of our two IAC sound attenuated chambers. These experimental rooms are also outfitted with two-channel earphones and are used for more exacting experimental paradigms such as psychoacoustic studies involving detection and discrimination, discrimination training and perceptual learning. One of the IAC booths is equipped as an off-line recording chamber and as a point of origin for natural speech data input directly to the computer's A/D system. These chambers are also fitted with 9" CRT monitors which can be used to interact with subjects in the same way as in the CRT laboratory room on the PDP 11/05 system. These monitors permit us to prompt a speaker who is providing natural speech to be input to the A/D system and provide a protocol for recording sessions involving a large number of stimulus tokens in speech production experiments.

Word Processing.

The same facility described above under the PDP 11/05 is also available on this system. The same two terminals used for word processing tasks are available via manual switch selection connected directly to the PDP 11/34 system when it is not being used for research activities.

Summary and Conclusions

It has been our intent in this report to present the reader with the design criteria, research goals and instrumentation strategies that have led to the development and implementation of the current computer facilities of the Speech Perception Laboratory. We view the current laboratory as just one step in a continual process of evolution; each completed research project giving birth to the next idea and its own unique set of instrumentation needs. The attributes of modularity, adaptability and functional transportability have been central to our thinking, planning and implementation of each laboratory facility. We have also placed high priority on developing a well organized and efficiently operating laboratory with readily available software support as we believe these are fundamental prerequisites for conducting high quality research and collecting reliable experimental data under highly controlled conditions. In addition, such a flexible laboratory facility permits maximum efficiency and use by a large number of researchers and students who are working on a wide diversity of problems in speech communication.

Footnotes

*The development and implementation of the computer facilities described in this report was supported by funds from NIMH research grant MH24027 and NINCDS research grant NS12179 to Indiana University in Bloomington.

References

Carrell, T. & Kewley-Port, D. Graphic Support for KLTEXC. RESEARCH ON SPEECH PERCEPTION Progress Report No. 4, Indiana University; 1978, 235-246.

Forshee, J. C. Speech Perception Research Laboratory: The State of the Computer System. RESEARCH ON SPEECH PERCEPTION Progress Report No. 2, Indiana University; 1975, 202-220.

Forshee, J. C. Computer Resources in the Speech Perception Laboratory. RESEARCH ON SPEECH PERCEPTION Progress Report No. 3, Indiana University; 1976, 185-201.

Kewley-Port, D. KLTEXC: Executive Program to Implement the KLATT Software Speech Synthesizer. RESEARCH ON SPEECH PERCEPTION Progress Report No. 4, Indiana University; 1978, 235-246.

Kewley-Port, D. SPECTRUM: A Program for Analyzing the Spectral Properties of Speech. RESEARCH ON SPEECH PERCEPTION Progress Report NO. 5, Indiana University; 1979, Pp. 475 - 492.

Sawusch, J. R. A description of the OVE IIId Control Program: OVEXEC. RESEARCH ON SPEECH PERCEPTION Progress Report No. 2, Indiana University; 1975, 221-226.

SPECTRUM: A Program for Analyzing the Spectral Properties of Speech

Diane Kewley-Port

This report describes the basic design of SPECTRUM, a FORTRAN program for digitally analyzing arious spectral properties of speech waveforms. SPECTRUM was developed for use in a small laboratory environment, where on-line, interactive computer graphics are basic to the computer processing techniques. The analytic algorithms are primarily those based on linear prediction analysis developed by Markel and Gray (1976). The report covers aspects of both program design and human engineering. It also give sample task protocols with examples of the graphic displays.

Introduction

The SPECTRUM program has been developed in the Speech Perception Laboratory to spectrally analyze speech waveforms. SPECTRUM was designed to permit flexible spectral analysis conditions using interactive computer graphics for user feedback. The primary analytic techniques were based on linear prediction analysis, but other types of routines have been implemented including discrete Fourier transforms and spectral analysis using a digital model of an auditory filter bank. To a large extent, SPECTRUM was patterned after the ILS (Interactive Laboratory Systems) programs developed at Speech Communication Research Laboratories, Los Angeles, Ca. The majority of the signal processing algorithms can be found in Markel and Gray's (1976) book, Linear Prediction of Speech, except for the FFT which was developed by Markel (1971). These algorithms, as well as all of SPECTRUM, were written in FORTRAN IV. SPECTRUM operates on both the PDP 11/05 and PDP 11/34 computers as described by Forshee (1979). Both machines have 28K of core and use the DEC RT-11 operating system.

Human Engineering

In developing a highly technical program like SPECTRUM, one must carfully consider the type of user and the most frequent program applications. In this case, we have users who are non-engineers and whose speech research interests often involve small data bases. The research applications usually involve questions concerning the spectral properties of speech which can be analyzed best using on-line, interactive graphics. It was therefore decided that a high priority goal was easy man-machine communication, even when this resulted occasionally in less flexible data manipulation.

One of the first decisions was to fix the analysis file structures and make them opaque to the user. Thus the user does not make decisions concerning the nature or contents of the output files, but on the other hand he may be saving more data than he needs. Operating SPECTRUM then requires both an input waveform file and an output analysis file to be defined at all times. SPECTRUM automatically saves almost all calculated values, such as linear prediction coefficients, formant values etc., in the integer-valued analysis file without the user specificaly designating which output values should be saved. Another rationale for this type of analysis file is that calculation time is relatively costly on our machines, at least at the present time, and the small data bases are often examined repeatedly in detail.

Another important decision was to base the analysis interval on the concept of the frame. The frame is defined by the number of waveform samples between each spectral analysis of a waveform segment. For example, to roughly calculate the formant tracks of a long utterance, a suitable analysis interval between formant values might be chosen as 25ms. The corresponding frame size for a sample rate of 10000 Hz would be 250 points. In order to time lock any given analysis file to a wavefrom file, the frame becomes the important defining characteristic of the analysis file. This gives the user an advantage of being able to analyze frames out of sequence, or recalculate frames using different analysis conditions. The user will always know absolutely what part of the waveform has been analyzed. Therefore, once the context has been set for a particular analysis file, it cannot be changed. However, several different analysis files having different contexts can be associated with a single waveform file.

The SPECTRUM commands are selected by means of two character names listed in the MENU shown in Table 1.

Insert Table 1 about here

After a command is accessed by SPECTRUM, a prompt message is displayed. In some cases another menu may also be displayed. These prompts contain mneumonices separated by commas to indicate the input arguments for the specific command. For example, command PP (i.e., pick the peaks from a spectral section) has the prompt:

STFR,NO.FR,ERASE->.

'STFR' means frame where the analysis is to start; 'NO.FR' means the number of frames included in the analysis; and 'ERASE' is a binary valued parameter to erase (or not erase) the graphics terminal before the PP ouput is displayed.

The structure of these prompts has been carefully designed. First, the prompts include all arguments used in the the command. The order of the arguments, as well as any abbreviations used, is preserved across commands. Second, input values for the arguments operate in a full default mode so that reasonable values for all arguments are assumed by SPECTRUM. (The default mode will be discussed further below.) Detailed documentation of command actions and arguments (similiar to the ILS format) are available to the user. Appendix A provides an example of the documentation for command SS.

Table 1

SPECTRUM MENU (August, 1979)

2 Character Command	Discription
AC	Compute analysis Coefficients
CS	Display Cursor
CT	Query or change Context
EX	EXit Spectrum
FT	Compute and display Formant Tracks
FR	Redisplay FT
GP	Query and change Global Parameters
LI	LIst on printer
MU	Menu display
OW	Open Waveform and analysis files
OA	Open new Analysis file
PP	Pick Peaks of spectrum and display
PR	Redisplay PP
QD	Query length display buffer
QP	Query Pointers
SI	SIft algorithm for pitch extraction
SW	Redisplay SI
SS	Calculate Smoothed Spectrum and display
SR	Redisplay SS
TD	Three-Dimensional spectrum display
TR	Redisplay TD
TI	TItle display
XF	Fast Fourier spectrum and display
XR	Redisplay XF
WA	WAveform display
WR	Redisplay WA

Program Structure

The program SPECTRUM consists of a root segment, a subroutine library and a set of command subroutines with essentially one subroutine per command. The root contains the FORTRAN COMMON blocks and the calling sequences for each command. The individual command subroutines are then used in overlay regions of core. Subroutines used by several commands, such as FFT, reside in the library.

The full default mode of the commands depends on passing argument values in COMMON. The two most important command arguments in COMMON are STFR and NO.FR. Once set by one command, they are retrieved from COMMON by all subsequent commands until they are reset. Another use of COMMON is to maintan the global analysis parameters which allow continuity of analysis condistions during processing. Parameters such as M (i.e., number of linear prediction coefficients), or NBITS (i.e., power of the FFT) are kept in a single global table. As global parameters are set for specific analyses, they can then be recalled in subsequent analyses from the COMMON. Furthermore, they are stored in the header of the analysis file so that the most recent global parameters are reset in COMMON anytime an analysis file is opened.

There are generally two types of output produced by a command. Typically, any values calculated are stored in the analysis file, and appropriate displays of these values are plotted on the graphics terminal (see Appendix A for examples of the SS displays). The listing command, LI, permits a listing of all stored values on either the line printer or the user's console. The displays are currently plotted on a DEC VT-11, a 17 inch, refresh-memory CRT. Displays are not usually flexible, but rather oriented to a particular task, as can be seen in the next section. Because the displays are constructed relatively slowly, the most recent display from a command is stored on disk for rapid retrieval (see the MENU). Currently hard-copy of the displays is obtained using a Poloroid CU-5 cammera attatched to a custom-made hood for the VT-11. All figures in this manuscript were originally photographed using this system. We expect delivery shortly of a Tektronics storage oscilloscope and hard copy unit to replace the camera system.

Sample Protocol with Displays

In order to illustrate the use of SPECTRUM, sample protocols for several research tasks will be outlined below. Please note that although the text is written as if one waveform file was being analyzed, the displays actually come from analyses of several different waveforms.

First the user may wish to analyze in detail the early formant transitions in a consonant-vowel syllable. A waveform file and analysis file are opened, and the frame size is set to 50 points or 5ms. The waveform is displayed for the first 30 frames (150ms) using command WA as shown in Figure 1.

Insert Figure 1 about here

The user may decide to analyze the first 22 frames, starting at Frame 1. The command AC calculates the linear prediction coefficients as inverse filter coefficients using the autocorrelation method. This analysis uses the subroutine AUTO from Markel and Gray (1976). Command PP is executed next to calculate both the smoothed spectral sections for each frame using subroutine FFTMGR from Markel, (1971), and to pick the spectral peaks for each frame using subroutine FINDPK from Markel & Gray (1976). The resulting display of the peaks and RMS energy for each frame is shown in Figure 2.

Insert Figure 2 about here

The formant tracks were calculated for the first four formants by FT, subroutine FORMNT fom Markel & Gray (1976). The display of the 22 frames of waveform, the peaks and the superimposed formants are shown in Figure 3.

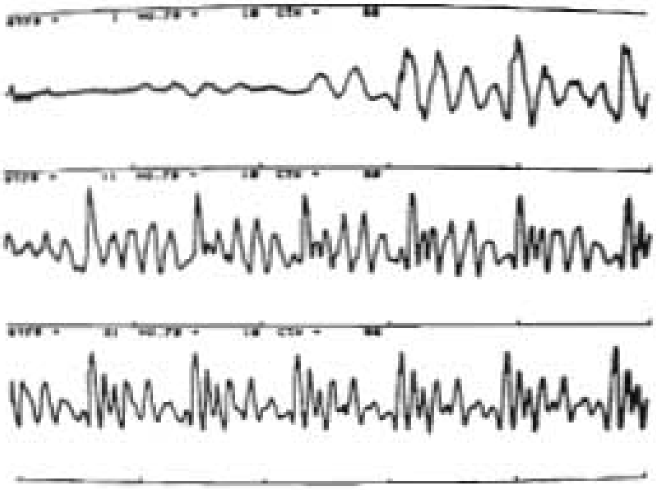


Figure 1. The first 30 frames of a waveform are displayed sequentially at three different locations on the CRT screen using command WA.

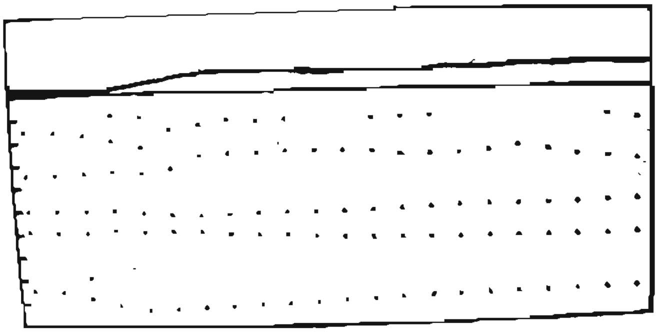


Figure 2. The display of the spectral peaks for each of 22 frames produced by the command PP, plus the RMS energy curve. The x-axis is time in frames. The y-axis ticks for the formant peaks are at 500 Hz intervals. The y-axis range for RMS energy is 0 to 2^{16}.

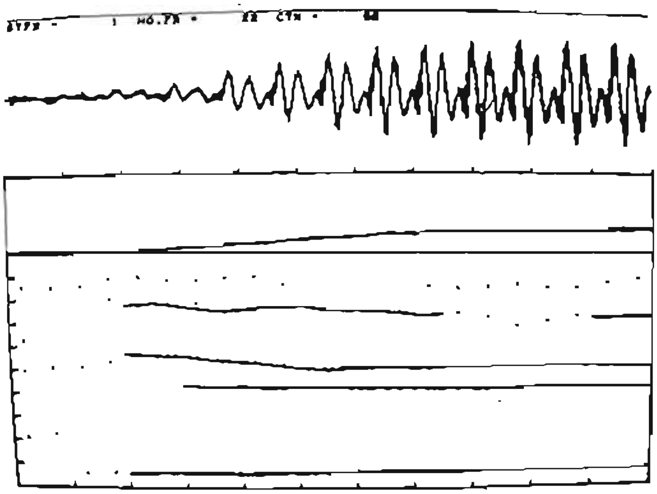


Figure 3. This figure shows the display of a waveform, the peaks of the spectral sections, the RMS energy and the formant tracks for 22 frames. The coordinates are the same as Figure 2, except for the presence of the x-axis ticks in 10ms intervals.

Insert Figure 3 about here

The user may then wish to examine the distribution of spectral energy over time in the first 8 frames of the waveform. Figure 4 shows a display of linear prediction smoothed spectra from command TD (three-dimensional display). Figure 5 also uses TD, but produces spectral sections more like the auditory filters described by Patterson, 1974. Figure 5 is titled using the command TI.

Insert Figures 4 and 5 about here

Finally, SPECTRUM can also be used to extract pitch from an utterance using the SIFT algorithm (Markel & Gray, 1976). Here the context is chosen as 250 points or 25ms which is more appropriate for pitch tracking since the rate of change in pitch occurs relatively slowly in speech compared to formant transitions. Again WA is used to select the appropriate portion of the waveform to analyze. The SIFT command SI calculates the pitch and displays it in the large format shown in Figure 6. A user can also obtain a listing of the numerical values of the fundamental frequency, or any other values calculated by SPECTRUM, using the listing command, LI.

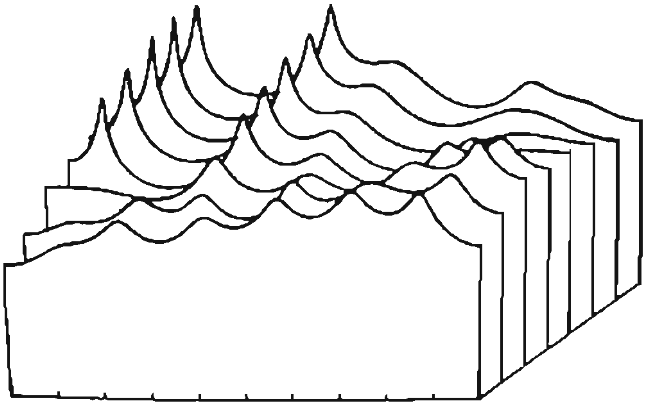


Figure 4. Three-dimensional display of the smoothed spectra of the inverse filter coefficients are plotted using TD. The x-axis is frequenccy in 500 Hz ticks. The y-axis is relative dB. The z-axis is in 5ms intervals which is the frame size for this analysis.

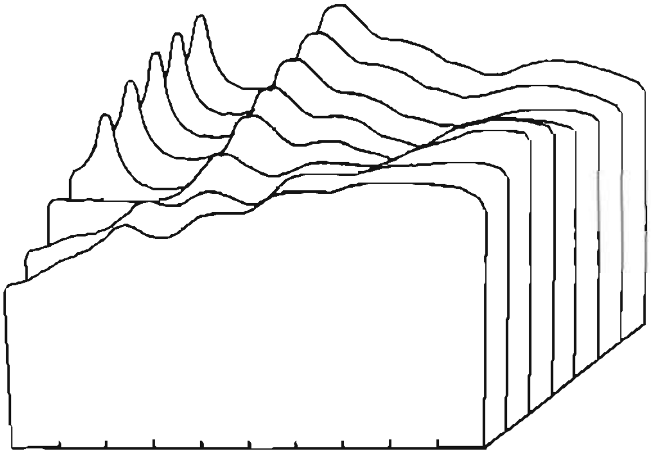


Figure 5. This display is almost the same as that of Figure 4, except that the waveform has been analyzed using routines to approximate filtering in the peripheral auditroy system.

Insert Figure 6 about here

Summary

This report has briefly described the spectral analysis program SPECTRUM as currently implemented in our laboratory. Since the program has many possible applications, it has grown and changed over the last year, and is expected to continue to change in the future as additional requirements and needs arise. Some of the possible additions to SPECTRUM are direct audio input and output capabilities, alternative methods for pitch extraction, and synthesis capabilities from the linear prediction coefficients. SPECTRUM is being used as a valuable tool for acoustic analysis of speech in the laboratory and is being used on a daily basis in speech production and synthesis research.

Acknowledgements

We are very grateful to Dr. June Shoup-Hummel and the Speech Communication Research Laboratory, Inc. Los Angeles for permitting us to obtain copies of The Interactive Laboratory System (ILS) programs and command manuals after which SPECTRUM was closely patterened. We are also grateful to Dr. Steven Davis for providing us with help and assistance with the ILS system during his stay at Haskins Laboratories. Dr. David Broad and Dr. Larry Pfeifer were also helpful in the early stages of this project. The

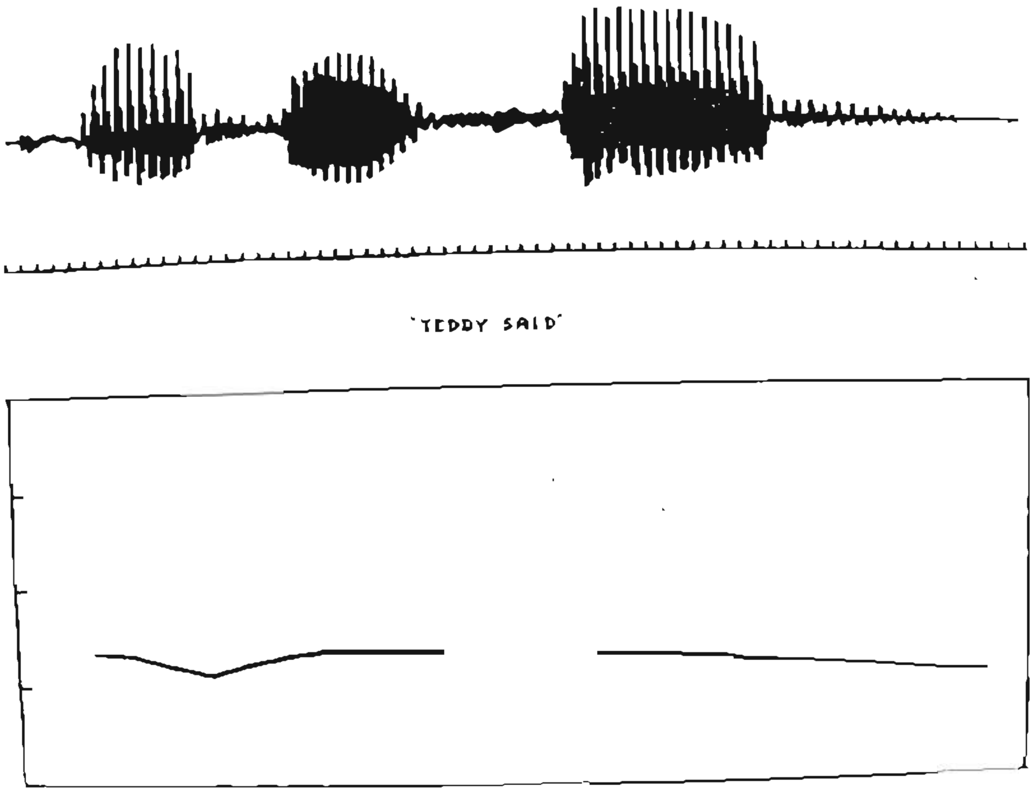


Figure 6. This display shows the pitch calculated by SI, aligned with the waveform. For the pitch axes, the x-axis is the same as for the waveform (marked in 10ms ticks), while the y-axis has ticks every 100 Hz. The utterance displayed is 'Teddy said'.

development of SPECTRUM was supported by NIH Grant NS-12179 to Indiana Univerisity.

References

Forshee, J. (1979) 'Computer facilities in the Speech Perception Laboratory.' RESEARCH IN SPEECH PERCEPTION Progress Report No. 5, Pp. 449-473.

Markel, J. D. (1971) 'FFT pruning'. IEEE Trans. AV-19, 305-311.

Markel, J. D. and A. H. Gray, jr. (1976) Linear prediction of Speech. Springer-Verlag, New York.

Patterson, R. D. (1974) 'Auditory filter shape.' J. Acous. Soc. Am. 55, 802-809.

SMOOTHED SPECTRUM COMMAND

****
SS
****

Function

To calculate the frequency spectrum of the inverse filter and display it for one frame. Can be compared to FFT of frame using XF or XR.

Prompt

STFR, ERASE, DB, LINE TYPE ->

Command Arguments

STFR = Frame to be displayed
If STFR = Ø, then default and value in COMMON used
If STFR > Ø, then COMMON value set to STFR.

ERASE = Erase screen before display
If ERASE = Ø, then default. Erase screen before display and save this display on disk for redisplay using SR.
If ERASE = other, keep current display. A total of 3 smoothed spectrum can be displayed simultaneously.

DB = dB value to add to spectrum to shift spectrum on display frame.
If DB = Ø, normalized gain spectrum displayed
If DB ≠ Ø, then checked that $-5\text{Ø} \leq \text{DB} \leq +5\text{Ø}$, and DB value added to each spectrum value before display.

LINE TYPE = spectrum drawn in line types offered in DECGRAPHICS
If LINE TYPE = Ø, then default value, and LINE TYPE set to 1, solid line.
If LINE TYPE = 2, display uses long-dashed line.
If LINE TYPE = 3, display uses short-dashed line.
If LINE TYPE = 4, display uses dot-dash lines.
If LINE TYPE > 4, error and legal value for line type requested.

Global Parameters Used

NBITS = Power of 2 of the FFT. Value of NBITS in GLOBAL COMMON is used to calculate FFT of analysis coefficients, and the resulting smoothed spectrum as stored in the analysis vector (see GP for further information).

Preparatory Commands

OW A waveform and an analysis file must be open.

AC Analysis coefficients of the inverse filter must be stored in the analysis vector for the frame specified.

Confirmatory Commands

Self-confirming in display

LI can be used to print the smoothed spectrum.

Description

The frequency spectrum of the inverse filter coefficients is a smoothed, amplitude-versus-frequency display of the model of the resonances of the vocal tract with the fundamental frequency removed. The SS command calculates the spectrum as log magnitude of the Fourier Transform in dB. The calculated spectrum in dB is stored in the analysis vector in dB and is displayed on the VT-11. The smooth spectrum is always gain normalized by adding the constant GAIN to the spectral values (see AC) and is further shifted by the constant DB when specified by the user. The smoothed spectrum can be compared to the Discrete Fourier Transform of the waveform using X or XR.

The SS program first reads the analysis vector for the specified frame (if it does not exist, SS aborts with a message). The analysis vector is checked to see whether the spectrum is already stored by previous use of SS or PP for this frame. If spectrum exists, the program proceeds to the display. If not, the analysis coefficients are read and NBITS is obtained from GLOBAL COMMON. A Fast Fourier Transform of the A coefficients is calculated so that the number of values of magnitude on output is 2**(NBITS -1). The magnitude is converted to dB for display and stored in the analysis vector as integer *512. 256 locations in the analysis vector are alloted to the spectrum, which limits NBITS to 9.

Examples

Fig. SS-1 shows the use of SS to contrast spectral sections of two adjacent frames. The commands were:

```
SS (ret)
STFR, ERASE, DB, LINE TYPE -> 3 (ret)
SS (ret)
STFR, ERASE, DB, LINE TYPE -> 4, 1 (ret)
```

The display of frame 3 was stored on the disk for rapid retrieval by SR because ERASE = Ø.

Fig. SS-2 demonstrates how the smoothed spectral sections produced by SS can be compared to an discrete Fourier Transform of nearly the same section of waveform using XF. The commands were:

```
SS (ret)
STFR, ERASE, DB, LINE TYPE -> 8 (ret)
XF (ret)
STFR, NPOINT, DB, LINE TYPE -> 8, 400, -20 (ret)
SR (ret)
ERASE -> (ret)
```

The first two commands displayed a spectral section for frame 4 and stored it on the disk. Using XF an 400 point FFT was calculated and displayed in the same display frame as SS, but offset by -20 dB. XF always erases the VT-11 screen first, so RR was used to retrieve the smoothed spectrum display.

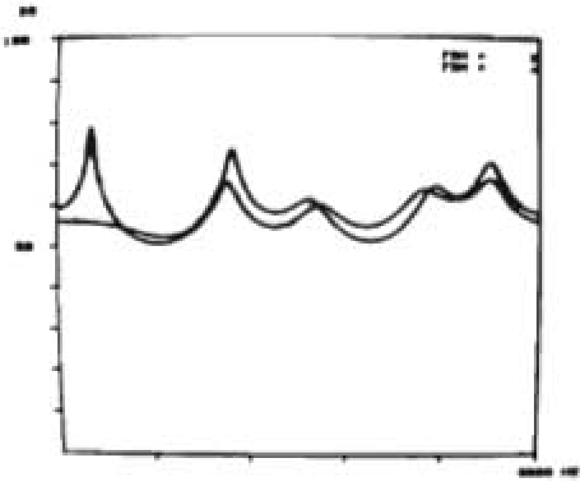


Fig. SS-1. Display showing two spectral sections produced by two calls to command SS.

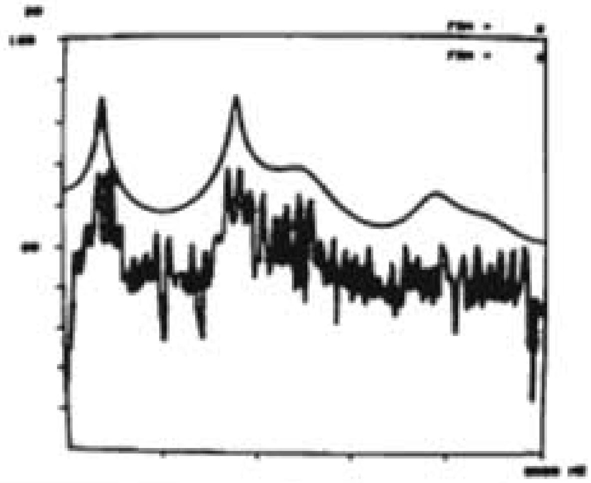


Fig. SS-2. Comparison of spectral section display produced by SS command and discrete Fourier Transform display produced by command XF.

IV. PUBLICATIONS

Publications:

Pisoni, D. B. Review of "Speech Recognition" by R. Reddy. *Journal of the Acoustical Society of America*, 1979, *65*, (3), 867-869.

Pisoni, D. B. Perception of Speech vs. Nonspeech: Evidence for Different Modes of Processing. *Proceedings of the Ninth International Congress of Phonetic Sciences*, Copenhagen, August, 1979. Copenhagen: Institute of Phonetics, University of Copenhagen, Pp. 433-437.

Grunke, M. E. & Pisoni, D. B. Perceptual Learning of Mirror-Image Acoustic Patterns. *Proceedings of the Ninth International Congress of Phonetic Sciences*, Copenhagen, August, 1979. Copenhagen: Institute of Phonetics, University of Copenhagen, Pp. 461-467.

Pisoni, D. B. On the perception of speech sounds as biologically significant signals. *Brain, Behavior, and Evolution*, 1979,

Pisoni, D. B. The Role of Early Experience in Sensory & Perceptual Development. In L. Harmon (Ed.), *Interrelations of the Communicative Senses*, Washington: National Science Foundation Final Report 1979, Pp. 296-307.

Bell-Berti, F., Raphael, L., Pisoni, D. B. and Sawusch, J. R. Some Relations between articulation and perception. *Phonetica*, 1979,

The following papers were presented at the Acoustical Society of America meeting in Cambridge, Mass., June 1979, and all appear in:

J. J. Wolf and D. H. Klatt (Eds.), *Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America*. New York: Acoustical Society of America, 1979.

Pisoni, D. B., Carrell, T. D. and Simnick, S. S. Does a listener need to Recover the Dynamic Vocal Tract Gestures of a Talker to Recognize His Vowels? Pp. 19-23.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L. and Perey, A. J. Identification and Discrimination of a New Linguistic Contrast. Pp. 439-442.

Carrell, T. D. and Smith, L. B. Some Perceptual Dependencies Between Vowel Color and Pitch. Pp. 33-35.

Kewley-Port, D. Spectral continuity of Burst and Formant Transitions as Cues to Place of Articulation in Stop Consonants. Pp. 175-178.

Gruenenfelder, T. M. Fundamental Frequency as a Cue to Postvocalic Consonantal Voicing: Some Data from Perception and Production, Pp. 57-61.

Remez, R. E. Adaptation of the category boundary between speech and nonspeech: A consonantal case against feature detectors. Pp. 337-340.

Grunke, M. E. Perceptual learning of acoustic patterns: Mirror-image vs. direction-of-glissando pairs. Pp. 451-454.

Manuscripts to be Published:

Aslin, R. N. & Pisoni, D. B. Some developmental processes in speech perception. In G. Yeni-Komshian, J. F. Kavanagh, & C. A. Ferguson (Eds.), *Child Phonology: Perception and Production*. New York: Academic Press, 1980 (In Press).

Aslin, R. N. & Pisoni, D. B. Effects of Early Linguistic Experience on Speech Discrimination by Infants: A Critique of Eilers, Gavin and Wilson (1979). *Child Development*, 1980 (In Press).

Pisoni, D. B. Adaptation of the Relative Onset Time of Two-Component Tones. *Perception & Psychophysics*, 1980 (In Press).

Jusczyk, P. W., Pisoni, D. B., Walley, A. and Murray, J. Discrimination of relative onset time on two-component tones by infants. *Journal of the Acoustical Society of America*, 1980 (In Press).

Pisoni, D. B. Some Measures of Intelligibility and Comprehension. In J. Allen (Ed.), *Conversion of Unrestricted English Text to Speech*. 1980 (In Press).

Remez, R. E. Susceptibility of a stop consonant to adaptation on a speech-nonspeech continuum: Further evidence against feature detectors in speech perception. *Perception & Psychophysics*, 1980 (In Press).

V. Laboratory Staff and Personnel:

David B. Pisoni, Ph.D.	Professor of Psychology
Richard N. Aslin, Ph.D.	Associate Professor of Psychology
Robert E. Remez, Ph.D.	Visiting Assistant Professor of Psychology

Michelle A. Blank, Ph.D.	NIH Post-doctoral Fellow
Mary Ellen Grunke, Ph.D.	NIH Post-doctoral Fellow
Sue Ellen Krause, Ph.D.	NIH Post-doctoral Fellow
Joan M. Sinnott, Ph.D.	NIH Post-doctoral Fellow

Jerry C. Forshee, M.A.	Computer Systems Analyst
Diane Kewley-Port, M.S.	Research Associate
Steven S. Simnick, M.A.	Research Assistant

Thomas D. Carrell, B.A.	Research Assistant
Wendy Crawford, B.A.	Research Assistant (Infant Laboratory)
Thomas Gruenenfelder, B.A.	Research Assistant
Beth L. Hennessy, B.A.	Research Assistant
Alan J. Perey, B.A.	Research Assistant
Amanda C. Walley, B.A.	Research Assistant

David Link	Electronics Engineer
Nancy Layman	Administrative Secretary

Special Consultants:

Peter W. Jusczyk, Ph.D.	Dalhousie University
James R. Sawusch, Ph.D.	SUNY at Buffalo