Published in final edited form as:
J Exp Psychol Learn Mem Cogn. 1988 July ; 14(3): 421–433.

Perceptual Learning of Synthetic Speech Produced by Rule

Steven L. Greenspan, **Howard C. Nusbaum**, and **David B. Pisoni**
Indiana University

Abstract

To examine the effects of stimulus structure and variability on perceptual learning, we compared transcription accuracy before and after training with synthetic speech produced by rule. Subjects were trained with either isolated words or fluent sentences of synthetic speech that were either novel stimuli or a fixed list of stimuli that was repeated. Subjects who were trained on the same stimuli every day improved as much as did the subjects who were given novel stimuli. In a second experiment, the size of the repeated stimulus set was reduced. Under these conditions, subjects trained with repeated stimuli did not generalize to novel stimuli as well as did subjects trained with novel stimuli. Our results suggest that perceptual learning depends on the degree to which the training stimuli characterize the underlying structure of the full stimulus set. Furthermore, we found that training with isolated words only increased the intelligibility of isolated words, although training with sentences increased the intelligibility of both isolated words and sentences.

Speech signals provide an especially interesting and important class of stimuli for studying the effect of stimulus variability on perceptual learning, primarily because of the lack of acoustic–phonetic invariance of the speech signal (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Despite large differences in the acoustic–phonetic structure of speech produced by different talkers, listeners seldom have any difficulty recognizing the speech produced by a novel talker. Although context-dependent and talker-dependent acoustic–phonetic variability has often been viewed as noise that must be stripped away from the speech signal in order to reveal invariant phonetic structures (e.g., Stevens & Blumstein, 1978), it is also possible that this variability serves as an important source of information for the listener, which indicates structural relations among acoustic cues as well as information about the talker (Elman & McClelland, 1986). If the sources of variability in the speech waveform are understood by the listener, this information may play an important role in the perceptual decoding of linguistic segments (see Liberman, 1970). Therefore, if a listener must learn to recognize speech that is either degraded or impoverished, information about acoustic–phonetic variability of the speech signal may be critical to the learning process.

Schwab, Nusbaum, and Pisoni (1985) recently demonstrated that moderate amounts of training with low-intelligibility synthetic speech will improve word recognition performance for novel stimuli generated by the same text-to-speech system. Schwab et al. trained subjects by presenting synthetic speech followed by immediate feedback in recognition tasks for words in isolation, for words in fluent meaningful sentences, and for words in fluent semantically anomalous sentences. Subjects trained under these conditions improved significantly in recognition performance for synthetic words in isolation and in sentence contexts compared to subjects who either received no training or received training on the

Correspondence concerning this article should be addressed to Howard C. Nusbaum, who is now at the Department of Behavioral Sciences, 5848 South University Avenue, The University of Chicago, Chicago, Illinois 60637..

same experimental tasks with natural speech. Thus, the improvement found for subjects trained with synthetic speech could not be ascribed to mere practice with or exposure to the test procedures. In addition, a follow-up study indicated that the effects of training with synthetic speech persisted even after 6 months. Thus, training with synthetic speech produced reliable and long-lasting improvements in the perception of words in isolation and of words in fluent sentences.

One interesting aspect of the study reported by Schwab et al. is that subjects were presented with novel words, sentences, and passages on every day of the experiment. Thus, these subjects were presented with a relatively large sample of synthetic speech during training, and as a result, these listeners perceptually sampled much of the structural variability in this "synthetic talker." The improvements in recognition of the synthetic speech may have been a direct result of learning the variability inherent in the acoustic–phonetic space of the text-to-speech system. On the other hand, listeners may simply have learned new prototypical acoustic–phonetic mappings (see Massaro & Oden, 1980) and ignored the structural relationships among these mappings.

Another interesting aspect of the Schwab et al. study is the finding that recognition improved both for words in isolation and for words in fluent speech. This finding is of some theoretical relevance because recognizing words in fluent speech presents a problem that is not present when words are presented in isolation: The context-conditioned variability between words and the lack of phonetic independence between adjacent acoustic segments leads to enormous problems for the segmentation of speech into psychologically meaningful units that can be used for recognition. In fluent, continuous speech it is extremely difficult to determine where one word ends and another begins if only acoustic criteria are used (Pisoni, 1985; although cf. Nakatani & Dukes, 1977).

Indeed, almost all current models and theories of auditory word recognition assume that word segmentation is a by-product of word recognition. Instead of proposing an explicit segmentation stage that generates word-length patterns that are matched against stored lexical representations in memory, current theories propose that words are recognized one at a time, in the sequence by which they are produced (Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986; Reddy, 1976). These theories claim that there is a lexical basis for segmentation such that recognition of the first word in an utterance determines the end of that word as well as the beginning of the next word. Although none of these models was proposed to address issues surrounding perceptual learning of words, these models suggest that training subjects with isolated words generated by a synthetic speech system should improve the recognition of words in fluent synthetic speech: If listeners recognize isolated words more accurately, word recognition in fluent speech should also improve, assuming that perception of words in fluent speech is a direct consequence of the same recognition processes that operate on isolated words. Conversely, training with fluent synthetic speech should improve performance on isolated words.

However, recent evidence from studies using visual stimuli suggests that differences in the perceived structure of training stimuli may lead to the acquisition of different types of perceptual skills. Kolers and Magee (1978) presented inverted printed text in a training task and instructed subjects either to name the individual letters in the text or to read the words. After extensive training, subjects were found to have improved only on the task for which they received training: Attending to letters improved performance with letters but had little effect on reading words; conversely, attending to words improved performance with words but had little effect on naming letters. However, results for visual stimuli may not necessarily apply to speech because of the substantial differences that exist between

spatially distributed, discrete printed text and temporally distributed, context-conditioned speech.

The present study was carried out to investigate the role of stimulus variability in perceptual learning and the operation of lexical segmentation as a consequence of word recognition. In the perceptual learning study carried out by Schwab et al. (1985), subjects were trained and tested on the same types of linguistic materials, but they were never presented with the same stimuli twice. In the present study, we manipulated the amount of stimulus variability presented to subjects during training and we trained different groups of subjects on different types of linguistic materials. Half of the subjects received novel stimuli on each training day, and the other half received a constant training set repeated over and over. The ability to generalize to novel stimuli should indicate how stimulus variability affects perceptual learning of speech. Also, half of the subjects were trained on isolated words, and half were trained on fluent sentences. Transfer of training from one set of materials to the other should indicate the effects of linguistic structure on perceptual learning.

Experiment 1

To examine the effects of training based on structural differences in linguistic materials, we trained subjects with either isolated words or with fluent sentences (but not both). The stimulus materials used in Experiment 1 were produced by the Votrax Type-'N-Talk text-to-speech system. A linear prediction coding analysis (Markel & Gray, 1976) of the Votrax-produced sentences indicated that a word excised from a fluent sentence was acoustically identical to the same word produced in isolation: Both words have identical formant structures and identical pitch and amplitude contours. Thus, the sentences produced by the Votrax system are merely end-to-end concatenations of individual words (with no pauses or coarticulation phenomena between words). The Votrax system does not introduce any systematic acoustic information in its fluent speech that is not already present in the production of isolated words. As a consequence, the Votrax-generated synthetic speech provides an excellent means of testing the claim that word segmentation is a direct consequence of word recognition. Because a sentence produced by the Votrax system is equivalent to a concatenated sequence of isolated words, improvements in recognizing isolated words should directly generalize to recognizing words in sentences.

In addition to examining the influence of the linguistic structure of training materials, we also investigated the effects of stimulus variability on perceptual learning. Some subjects received novel words or sentences throughout training, so that they never heard any stimulus more than once. Other subjects received a fixed set of words or sentences that was repeated several times during training. Both groups were tested on novel stimuli before and after training to examine generalization learning to novel words and sentences. Based on earlier research in pattern learning (e.g., Dukes & Bevan, 1967; Posner & Keele, 1968), subjects trained on novel exemplars should show more improvement than do those trained with repeated exemplars, because the novel training set provides a broader sample of the acoustic-phonetic variability in Votrax speech than is provided by the repeated training set. However, from artificial-language learning studies (e.g., Nagata, 1976; Palermo & Parish, 1971), we predict that if the repeated training set sufficiently characterizes the underlying acoustic–phonetic structure of the synthetic speech, there should be no performance difference between subjects receiving repeated training stimuli and those presented with novel training stimuli (as long as both sets of subjects receive an equal number of exemplars).

Method

Subjects—Sixty-six subjects participated in this experiment. All were students at Indiana University and were paid four dollars for each day of the experiment. All subjects were native speakers of English and reported having had no previous exposure to synthetic speech and no history of a hearing or speech disorder. All subjects were right-handed and were recruited from a paid subject pool maintained by the Speech Research Laboratory of Indiana University.

Materials—All stimuli were produced at a natural-sounding speech rate by a Votrax Type-'N-Talk text-to-speech system controlled by a PDP-11 computer. The Votrax system was chosen for generating words and sentences because of the relatively poor quality of its segmental (i.e., consonant and vowel) synthesis. Thus, the likelihood of ceiling effects in word recognition were minimized. The synthetic speech stimuli used in the present study were a subset of the stimuli developed and used by Schwab et al. (1985) to ensure comparability between experiments.

All stimulus materials were initially recorded on audiotape. After the audio recordings were made, the stimulus materials were sampled at 10 kHz, low-pass filtered at 4.8 kHz, digitized through a 12-bit A/D converter, and stored in digital form on a disk with the PDP-11/34 computer. The stimuli were presented in real time at 10 kHz through a 12-bit D/A converter and low-pass filtered at 4.8 kHz. Four sets of stimulus materials were used in this experiment:

1. *PB lists*. These stimuli consisted of 12 lists of 50 monosyllabic, phonetically balanced (PB) words (Egan, 1948).
2. *MRT lists*. These materials consisted of four lists of 50 monosyllabic consonant-vowel-consonant words taken from the Modified Rhyme Test (MRT) developed by House, Williams, Hecker, and Kryter (1965).
3. *Harvard sentences*. These stimuli consisted of 10 lists of 10 Harvard psychoacoustic sentences (IEEE, 1969; Egan, 1948). These are meaningful English sentences containing five key words plus a variable number of function words. The key words all contained one or two syllables.
4. *Haskins sentences*. These materials consisted of 10 lists of 10 syntactically normal, but semantically anomalous sentences that had been developed at Haskins Laboratories (Nye & Gaitenby, 1974). Each Haskins sentence contains four high-frequency monosyllabic key words presented in the following syntactic structure: “The (*adjective*) (*noun*) (*verb, past tense*) the (*noun*).”

Design—The entire experiment was conducted in six 1-hr sessions on different days. Five groups of subjects were tested on the first and last day of the experiment. The 4 intervening days were used to provide training for subjects in four of the five groups. A weekend separated the pretraining test session (on Day 1) from the first day of training (Day 2). All groups were treated similarly during the pretraining and posttraining test sessions (Days 1 and 6). For each test, subjects received the same materials in the same order. However, different materials were presented on the two testing days.

Each group was treated differently during training. One group of subjects (the novel-word group) received a different set of isolated words on each day of training, whereas a second group (the novel-sentence group) received a different set of sentences each day of training. A third group (the repeated-word group) received a set of isolated words on the first day of training that was repeated on every subsequent day. Similarly, a fourth group of subjects

(the repeated-sentence group) received a set of fluent sentences on the first day of training that was repeated for all other training sessions. Thus, in the repeated-stimulus conditions, subjects were presented with different sets of stimuli only on the two test days (Days 1 and 6) and on the first day of training. The last group of subjects (the control group) received no training and provided a baseline against which the performance of the other four groups could be compared.

Procedure—All stimuli were presented to subjects in real time under computer control through matched and calibrated TDH-39 headphones at 77 dB SPL measured using an RMS voltmeter. Before each experimental session, signal amplitudes were calibrated using the same isolated word (from a PB list). Subjects were tested in groups with a maximum of 6 subjects per group.

Testing procedure—All subjects were tested before and after training, on Days 1 and 6 of the experiment, using the same test procedures, the same order of tasks, and the same sets of stimuli. It is important to note that although each group was tested on the same stimuli, the stimuli presented on Day 1 were different from those presented on Day 6. Moreover, the stimuli presented during these two test sessions were not used in any of the training sessions. Each test session lasted about 1 hr. No feedback was presented in any of the test-session tasks. In each test session subjects listened to 100 PB words, 100 MRT words, 10 Harvard sentences, and 10 Haskins sentences, in that order.

1. *PB task*. Subjects listened to 100 monosyllabic words, one word per trial in two blocks of 50 trials each. Each presentation of a spoken word was preceded by a 1-s warning light. The spoken word was presented 500 ms after the offset of the warning light. After each word was presented, subjects were asked to write the English word that they had heard on a numbered line in their response booklets. They were instructed to guess if they could not identify the word. Subjects were instructed to press a button on a computer-controlled response box when they completed writing their response for each trial. The next trial began 250 ms after all subjects had completed writing their responses or after 8 s had elapsed. Identification accuracy was scored from the written responses.

2. *Modified rhyme task*. In the MRT, subjects listened to 100 consonant-vowel-consonant words, presented one word per trial. At the start of each trial, a 1-s warning light was presented, followed by a 500-ms waiting period, which was followed by the spoken word. Immediately following the spoken word, six words were visually presented in a horizontal row centered on a CRT. Subjects were instructed to identify the spoken word by choosing a response from the six alternative words. Responses were indicated by pressing the corresponding button on a six-button computer-controlled response box. Responses were recorded and scored automatically by computer. The subjects were instructed to respond as accurately as possible. The next trial began 250 ms after all of the subjects responded or after an 8-s interval elapsed.

3. *Harvard sentence task*. Subjects listened to 10 sentences, presented one sentence per trial. The presentation of each sentence began 500 ms after the offset of a 1 -s warning light. Following each sentence, subjects were allowed up to 25 s to transcribe the sentence in a response booklet. Subjects were told to guess if they could not identify a particular word. Each page of the Harvard-sentence response booklet contained a column of 10 numbered lines, each of which was long enough for a response. The next trial began 250 ms after all the subjects pressed a button on their response boxes to indicate that they had finished responding or after 8 s had elapsed. Identification accuracy was scored from the written responses.

4. *Haskins sentence task*. Ten Haskins sentences were presented, one per trial, following the same procedure used for Harvard sentences. However, because each Haskins sentence has the same syntactic structure, the Haskins-sentence response booklet was constructed to reflect this structure. Each numbered line of the booklet was of the form: "The_____ _____ the _____"; subjects were instructed to write one word in each blank. Identification accuracy was scored from the written responses.

Training procedures—Except for the control group, all subjects received training with synthetic speech on Days 2 through 5 of the experiment. As in the testing sessions, on each trial the spoken stimulus was preceded by a 1-s warning light. After listening to a single spoken stimulus, subjects wrote the word or sentence in a response booklet. Half of the subjects receiving training listened only to isolated words (100 PB words per training session in two 50-word blocks), and half listened only to sentences (10 Harvard and 10 Haskins sentences).

After all of the subjects indicated that they had finished writing their response by pressing a button on a computer-controlled response box, feedback was provided: The correct word or sentence was displayed on a CRT in front of each subject and repeated once over the subject's headphones. The onset of the visual and auditory events were simultaneous. Subjects were instructed to press a button on a computer-controlled response box to indicate if they had correctly identified the stimulus for that trial. The visual information was displayed until all of the subjects had responded in this way or until 3 s had elapsed. The next trial began 250 ms after all subjects had pressed either the "correct" or "incorrect" response button. Except for the presentation of feedback, the procedure used in each training task (i.e., PB, Harvard, or Haskins) was identical to the procedure used in that task during the test sessions. During training sessions, the MRT was not presented.

On each day of training, novel-word subjects were presented with 100 new PB words, whereas the novel-sentence subjects heard 20 new Harvard sentences and 20 new Haskins sentences. On the first day of training, the repeated-word subjects were presented with the same 100 PB words that were presented to the novel-word subjects on the first day of training. However, for the repeated-word subjects, this same list was presented again on each subsequent training day. Similarly, the repeated-sentence subjects were first presented with the same 40 sentences that were presented to the novel-sentence subjects on their first day of training, and this list of sentences was presented again on each subsequent day of training.

Results

Six subjects did not complete the experiment, and their data were excluded from statistical analyses. Of the remaining 60 subjects, 12 subjects received novel-word training, 12 received novel-sentence training, 13 received repeated-word training, 11 received repeated-sentence training, and 12 received no training.

To score a correct response in the PB, Harvard, and Haskins tasks, subjects had to transcribe the exact word (or homonym) with no additional or missing phonemes. For example, if the word was *flew* and the response was *flute* or *few*, the response was scored as incorrect. However, *flue* would have been scored as a correct response. The results are reported separately for the isolated word and sentence tasks.

Isolated word recognition—Mean percentage of correct performance on the PB task and the MRT for the five groups is presented in Table 1 for the pre- and posttraining tests. There were significant differences in performance among the groups as a result of different types

of training: $F(4, 50) = 3.52$, $MS_e = 0.00346$, $p < .01$, for the MRT; $F(4, 55) = 2.87$, $MS_e = 0.00444$, $p < .05$, for PB words. Also, performance improved significantly from the pretraining test to the posttraining test for both tasks: $F(1, 50) = 97.8$, $MS_e = 0.00159$, $p < .01$, for the MRT; $F(1, 55) = 546.0$, $MS_e = 0.00215$, $p < .01$, for the PB words. Furthermore, the degree of improvement varied as a function of the type of training received by different subject groups: $F(4, 50) = 6.03$, $MS_e = 0.00159$, $p < .01$, for the MRT; $F(4, 55) = 10.04$, $MS_e = 0.00215$, $p < .01$, for the PB words.

A Newman-Keuls analysis of the pretraining scores indicated no reliable differences in performance among the groups prior to training for either task. However, an examination of the improvement scores (i.e., the difference between pre- and posttraining test performance) revealed that all of the groups receiving training improved significantly from the pretraining to the posttraining session for both isolated word tasks ($p < .01$). In contrast, the control group did not demonstrate any reliable evidence of improvement for the MRT or PB words. Moreover, a Newman-Keuls analysis of the improvement scores indicated that each of the training groups differed significantly from the control group for both isolated word tasks ($p < .05$) but not from each other. These results clearly demonstrate that training with either isolated words or sentences produces equivalent improvements in performance. Furthermore, these results suggest that, as long as the number of stimulus presentations are equivalent, training with a repeated set of stimuli produces as much improvement as does training with novel stimuli.

Sentences—Each sentence contained either four (Haskins sentences) or five (Harvard sentences) key words that were scored for recognition accuracy. Table 2 displays the average percentage of accuracy scores for each group of subjects in the pretraining and posttraining test sessions for two sentence tasks. No overall significant change in performance was observed from the pretraining to the posttraining test session: $F(1, 55) = 1.7$, $MS_e = 0.0068$, $p > .10$, for the Harvard sentences, whereas a significant improvement occurred for the Haskins sentences, $F(1, 55) = 538.0$, $MS_e = 0.00465$, $p < .01$. However, for both sets of sentences there was a significant effect of type of training on performance: $F(4, 55) = 7.95$, $MS_e = 0.0124$, $p < .01$, for Harvard sentences; $F(4, 55) = 15.8$, $MS_e = 0.01004$, $p < .01$, for Haskins sentences. Moreover, different types of training produced different amounts of improvement: $F(4, 55) = 14.93$, $MS_e = 0.0068$, $p < .01$; for Harvard sentences; $F(4, 55) = 17.2$, $MS_e = 0.00465$, $p < .01$, for Haskins sentences.

Planned comparisons indicated that mean performance for the control, word-trained, and sentence-trained groups did not differ significantly on the pretraining test (Day 1) for either set of sentences. However, 5 days later, word recognition in the Harvard and Haskins sentences was significantly different for the different subject groups, and the patterns of performance were different for the two types of sentences. For the Harvard sentences, planned comparisons indicated that performance dropped significantly for control subjects and for subjects who were trained with novel words ($p < .05$). However, subjects trained with repeated words showed no reliable increase or decrease in performance. In contrast, subjects trained with either novel or repeated sentences demonstrated a significant improvement in word recognition accuracy ($p < .01$). A Newman-Keuls analysis of the improvement scores demonstrated that the novel- and repeated-sentence conditions each produced significantly greater improvement scores than did the repeated-word, novel-word, and control conditions ($p < .01$). It is not obvious why the performance of the control and novel-word trained subjects decreased from the pretraining to the posttraining test.

One possible account of this decrease is that the Harvard sentences used in the posttraining test were significantly more difficult that those used in the pretraining test. There is some support for this materials-effect account because one of the groups in the study reported by

Schwab et al. (1985) showed a similar performance decrement for these sentences. However, an argument against this account is the fact that the repeated-word subjects in our study and the control group in the Schwab et al. study did *not* show this decrement. Of course, the most important aspect of these data is that the subjects who received training on these sentences improved in recognition after training.

For the Haskins sentences, planned comparisons indicated that all subjects displayed significantly better performance after training ($p < .01$). A Newman-Keuls analysis indicated that the improvement shown by the novel- and repeated-sentence groups was not significantly different from each other but was significantly greater than the control, novel-word, and repeated-word groups ($p < .01$). Moreover, the control, novel-word, and repeated-word groups did not differ reliably after training.

Moreover, the improvement demonstrated by the novel-word and repeated-word groups did not differ reliably from the improvement shown by the control group. The lack of a significant difference between these conditions could have arisen because training on isolated words has no effect on learning to recognize words in sentences, or because the experimental design was not powerful enough to detect differences between these conditions.

One method of increasing the power of the design is to derive scores for word training and for sentence training by averaging across the novel and repeated conditions and the two sentence tasks (i.e., the Harvard and Haskins tasks). The resulting improvement scores (accuracy percentage on Day 6 minus accuracy percentage on Day 1) were 3.6% for the control condition, 8.4% for the word-trained condition, and 28.4% for the sentence-trained condition. According to the Newman-Keuls multiple-range statistic, there was no reliable difference in improvement between the word-trained and control subjects, $p > .10$, but the differences between the sentence-trained subjects and the other groups were significant, $p < .01$. Therefore, there is no evidence in the present study that training on isolated words produces a reliable improvement in recognizing words in sentences. Of course, this is a null result and should be accepted subject to the usual cautions.

However, it is clear that improvement of performance on the sentence tasks is significantly more pronounced when subjects were trained with sentences than when they were only trained with isolated words. By comparison, performance on isolated word tasks increased significantly when subjects were given either sentence-training or word-training. Moreover, for the isolated-word tasks, the improvement demonstrated by the sentence-trained subjects was not reliably different from that of the word-trained subjects. Clearly, sentence training and word training are equally advantageous for isolated word recognition but not for identifying words in fluent speech.

Taken together, the results from the isolated word tasks and sentence tasks presented in the pre- and posttraining tests show that (a) all subjects who received training demonstrated similar improvements in recognition of isolated, novel words; (b) subjects trained on isolated words did not demonstrate any improvement in recognizing words in sentences, as compared to control subjects, whereas subjects trained on sentences improved significantly on recognizing words in novel sentences; and (c) there were no observable differences in the effects of training with novel or repeated stimuli.

One account of the difference in performance between word-trained and sentence-trained subjects is that subjects trained with isolated words may have difficulty locating the beginnings and endings of words in Votrax sentences. In isolated-word recognition tasks, the beginning and ending of each word is clearly marked by a period of silence. However, connected Votrax speech does not contain any physical segmentation cues to provide the

listener with an indication of the location of word boundaries. If explicit acoustic word boundaries aid in word recognition by segmenting fluent natural speech into word-size auditory units, subjects trained with isolated words might have expected and needed these boundaries to recognize the words in sentences produced by the Votrax system. In contrast, subjects trained with sentences of synthetic speech may have learned explicitly to recognize words without prior segmentation and may have developed a strategy of segmenting words by recognizing them one at a time in the order by which they are produced. This strategy would provide a recognition advantage for the sentence-trained subjects compared to subjects trained with isolated words. This account also suggests that the performance of word-trained subjects might improve when word boundary cues are available. This hypothesis was evaluated by a more detailed analysis of the recognition performance in the Haskins-sentence task.

Each Haskins sentence was constructed with a fixed syntactic frame such that each sentence contained two key words following the definite article *the* and two key words following open-class items (i.e., an adjective, noun, or verb) that varied from sentence to sentence. Subjects were told explicitly about this invariant syntactic structure and the response sheets displayed this structure for each sentence, with separate blanks for each open-class item and the word “the” for each occurrence of the definite article. If word-trained subjects had difficulty locating the beginnings of words, the recognition performance of these subjects, compared to control subjects, might be better for words following the definite article, because the relative locations of the definite articles were known in advance and the word-trained subjects would have better isolated-word recognition skills. In contrast, the performance of the word-trained and control groups should not differ on the words following an open-class item because no word boundary cues were provided for these items on the response sheets.

In order to evaluate the hypothesis that word-trained subjects recognized words following “the” more accurately than words following an open-class item, the data from the Haskins-sentence task were reanalyzed to include word position (words following “the” or an open-class item) as a factor. The means for the different treatment combinations are shown in Table 3 along with the amount of improvement that resulted from training.

After training, recognition performance on words that immediately followed the definite article was significantly better than was performance on words that followed an open-class item (67% vs. 36%), $F(1, 55) = 278.0$, $MS_e = 0.00970$, $p < .01$. Also, there was a significant interaction between word position and type of training, $F(4, 55) = 6.78$, $MS_e = 0.0097$, $p < .01$. An examination of the improvement scores (shown in Table 3) reveals that much of the improvement demonstrated by the control group and the word-trained groups was due to increased recognition performance on words following a definite article.

Newman-Keuls analyses revealed that for words following an open-class item, sentence-trained subjects improved significantly more than did word-trained and control subjects ($p < .01$). Moreover, the improvement demonstrated by word-trained subjects' scores did not reliably differ from the improvement shown by control subjects. The pattern of results for words following the definite article was quite different. The improvement scores for word-trained subjects were significantly higher than were those for the control subjects ($p < .05$) but significantly lower than were those of the sentence-trained subjects ($p < .05$). (The difference between sentence-trained and control subjects was also significant, $p < .01$.) As predicted, prior experience with isolated words aided recognition of words in sentences only when the identity of the preceding word was known in advance, which provided a cue to the location of a word's beginning. Subjects trained with isolated words were able to recognize words in sentences more accurately than were control subjects when some location

information was provided. Thus, the isolated-word training only improved recognition of isolated or segmented word patterns.

These data from the anomalous sentences suggest that word-trained subjects were not able to separate their perception of the acoustic–phonetic structure of a preceding word from that of a subsequent word, except when they had prior, reliable information about the identity of the preceding word (see Nakatani & Dukes, 1977; Nusbaum & Pisoni, 1985). With that one exception, training listeners to recognize words in sentences required specific experience with connected speech. These results demonstrate that improvements in recognizing isolated words do not necessarily predict improvements in recognizing words in sentences. Thus, the present findings argue against the hypothesis that word segmentation is a direct result of word recognition. In fact, it is possible that the skill that sentence-trained subjects acquired over and above the skills acquired by the word-trained subjects may involve explicitly learning the strategy of segmenting fluent speech via word recognition.

Although word-trained subjects were not able to generalize their newly acquired skills to recognizing words in fluent sentences, sentence-trained subjects did improve at recognizing isolated words. This finding suggests that the perceptual skills acquired in learning to recognize words in fluent sentences form a functional superset of the skills acquired during training with isolated words. It is interesting to compare this finding with those reported by Kolers and Magee (1978). In the Kolers and Magee study, training subjects to recognize inverted letters did not transfer to their reading inverted words, and training with inverted words had little impact on naming inverted letters. Thus, Kolers and Magee found little evidence for transfer even though visual words are structurally a superset of letters (just as spoken sentences are composed of words). The difference in the studies may arise from the differences between auditory and visual modalities. Printed letters are discrete stimuli, segmented from one another by blank spaces; in contrast, there are no silent intervals separating spoken words from their neighbors. Moreover, coarticulation effects often span word boundaries. One implication is that conclusions about perceptual learning of orthography cannot be generally applied, without great caution, to the domain of speech perception, and vice versa (see Liberman et al., 1967).

Training data—Although there were no systematic differences in the effects of training with novel or repeated stimuli on posttraining test performance, the day-by-day training data reveal differences between these types of training. These data show systematic improvements as a result of novel and repeated stimulus training with isolated words (Figure 1, top panel), Harvard sentences (Figure 1, middle panel), and Haskins sentences (Figure 1, bottom panel). For all three types of stimulus materials, recognition performance of the repeated-stimulus groups was significantly higher than was performance of the novel stimulus groups: PB words, $F(1, 23) = 49.6$, $MS_e = 0.0129$, $p < .01$; Harvard sentences, $F(1, 21) = 52.6$, $MS_e = 0.0202$, $p < .01$; and Haskins sentences, $F(1, 21) = 17.2$, $MS_e = 0.0195$, $p < .01$. In addition, overall recognition performance improved significantly on each day of training for all stimulus materials: PB words, $F(3, 69) = 279.7$, $MS_e = 0.0018$, $p < .01$; Harvard sentences, $F(3, 63) = 214.6$, $MS_e = 0.00228$, $p < .01$; and Haskins sentences, $F(3, 63) = 141.6$, $MS_e = 0.00233$, $p < .01$.

However, of greatest interest is the finding that the amount of learning in each training session depended on whether the training was based on repeated or novel stimuli: PB words, $F(3, 69) = 35.5$, $MS_e = 0.0018$, $p < .01$; Harvard sentences, $F(3, 63) = 34.3$, $MS_e = 0.00228$, $p < .01$; and Haskins sentences, $F(3, 63) = 11.6$, $MS_e = 0.00233$, $p < .01$. For all three types of stimulus materials, paired comparisons indicated that the interaction in performance between repeated-versus novel-stimulus training sessions was due to the absence of any significant difference between the two types of training on the first day of training (Day 2),

followed by significantly better recognition performance for repetition training for all subsequent days of training ($p < .01$). Furthermore, although it is not surprising that subjects trained with novel stimuli continued to show systematic improvements in performance over the training period, it is interesting that subjects trained on repeated stimuli also continued to improve in recognizing these stimuli throughout training ($p < .05$). The fact that subjects trained with repeated stimuli did not reach asymptotic performance after a single training session and continued to improve throughout training indicates that the structural complexity of the repeated stimuli could not be completely learned in a short period of time.

Despite variations in type of task (closed-response set procedures vs. open-response set procedures), type of stimulus materials (sentences vs. words), and type of training (sentences vs. words), no significant differences were observed in perceptual learning based on novel versus repeated training sets. One account of the present results is based on the observation that the variations in synthetic speech that must be learned are lawful and rule-governed much as the variations in artificial-language materials are (e.g., Nagata, 1976; Palermo & Parish, 1971). Under these conditions, learning of a repeated training set appears to produce the same level of generalization as does training with novel stimuli as long as the number of presentations is the same for the two training procedures. However, the utility of the repeated training set for generalization learning may depend on the degree to which the repeated stimuli characterize the underlying structure of the entire ensemble of possible stimuli (Palermo & Parish, 1971). In this context, it is useful to recall that Posner and Keele (1968) found that subjects trained with a variable set of highly distorted exemplars classified new, distorted exemplars more accurately than did subjects trained with a less variable, less distorted set. Considered together, these studies suggest that the structural relation between the training and test stimuli is far more important for perceptual learning than is simply the relative number of novel stimuli presented during training.

Discussion

Considering current theories of auditory word recognition, our findings are somewhat surprising. Almost all of these theories predict that improvements in recognition of isolated words should result in improved recognition of words in sentences. This prediction should be especially true for speech generated by the Votrax system because fluent connected speech produced by this device consists of concatenated strings of isolated words: There are no sentence-level phenomena in this synthetic speech. Our findings indicate that sentence-trained subjects learned something about sentences that could not be learned from training with isolated words alone. One hypothesis is that sentence-trained subjects learned to recognize words in sentences without *explicit* segmentation cues. A corollary to this hypothesis is that word-trained subjects performed poorly at recognizing words in sentences because they were unable to use the type of segmentation processes that normally would be used with natural speech. The inability to use these procedures dictates the need for learning a strategy that most theorists of word recognition attribute to the listener as part of normative word recognition—segmentation by recognition. Perhaps it is this strategy that was learned by subjects trained with fluent sentences.

Another major finding of the present study was that subjects trained with the same stimuli every day showed as much generalization learning as did subjects trained with novel stimuli. These results suggest that generalization learning is not dependent on training with novel stimuli. This is, in some respects, a surprising finding because subjects trained with novel stimuli every day were continually engaged in generalization. Subjects trained with repeated stimuli did not engage in generalization until the final session of the study. As a consequence, we might expect subjects with more experience at generalization to perform better on a generalization task, whereas subjects trained on a fixed set of stimuli might

perform much better on those stimuli but show little generalization to completely new stimuli.

One explanation for this outcome may be found in the training data. In general, repeated-stimulus and novel-stimulus groups continued to improve in performance throughout the training sessions without reaching an asymptote in accuracy. Thus, it is clear that subjects did not quickly or easily master the training set even though it was presented on each day with feedback. The apparent difficulty in learning this repeated training set may reflect the degree to which the training set characterizes the rules used by the Votrax in synthesizing speech. As in previous research on artificial-language learning, if the training set sufficiently describes the actions of the rules, generalization learning can occur even if the training set is relatively restricted.

Alternatively, the equivalent effectiveness of training with repeated and novel stimuli simply could be due to the number of training stimuli presented rather than to the structural complexity of the training set. Given that the subjects in repeated- and novel-stimulus training groups had the same exposure to synthetic speech and the same amount of feedback, it is possible that generalization learning in this experiment was due strictly to the number of stimulus presentations during training for each group. In Experiment 2 we tested this hypothesis by presenting two groups of subjects with the same number of stimuli during training; however, we substantially reduced the structural complexity of the training set for one of the groups.

Experiment 2

From the results of Experiment 1, it is tempting to conclude that the set of stimuli presented in the repeated stimulus condition was complex and varied enough to provide a reasonable characterization of the underlying structure of the entire ensemble of synthetic speech generated by the Votrax text-to-speech system. Learning the acoustic–phonetic structure of the synthetic speech may have been aided by prior knowledge of the acoustic–phonetic and lexical structures of English. Thus, although subjects in previous studies learned highly arbitrary and novel stimulus-response mappings, the subjects in Experiment 1 learned to map a “distorted” set of acoustic–phonetic cues onto a previously well-learned set of relations among natural acoustic–phonetic cues and lexical knowledge. Subjects may have modified existing knowledge structures to incorporate new acoustic–phonetic representations.

An alternative account of the perceptual learning observed in the first experiment is that simple exposure to the mechanical “sound” or voice quality of the Votrax-generated speech may improve the intelligibility of the speech. According to this hypothesis, there should be no difference between repeated and novel stimulus conditions as long as the amount of exposure to the synthetic speech is the same in repeated-and novel-stimulus training conditions. To investigate this hypothesis, we trained subjects on 200 novel PB words or on a fixed set of 10 PB words that was repeated 20 times. This small set of repeated stimuli is unlikely to provide a reasonable characterization of the underlying acoustic–phonetic properties of the text-to-speech system and is also very likely to be learned completely with a small number of repetitions.

Method

Subjects—Seventy-two undergraduates participated in this experiment to fulfill a requirement for an introductory psychology course. All subjects reported that English was their first language, that they had no history of hearing or speech disorders, and that they had had no prior exposure to synthetic speech.

Design—Subjects were assigned to two groups, each with 36 participants. Both groups were given a pretraining, open-response set test of 50 PB words at the beginning of a 1-hr session, and both received a different posttraining open-response set test of 50 PB words at the end of the session. However, in the interval between the pretraining and posttraining tests, the two groups were trained with different types of materials. One group was trained with 200 novel PB words divided into four blocks. The other group received a fixed list of 10 PB words that was repeated 20 times during training.

Materials—The PB word lists used in Experiment 1 were modified for this experiment. Four of the eight 50-word lists used for training in Experiment 1 were used without modification for training subjects in the novel-word condition. The list of 10 words that was used to train subjects in the repeated-word condition was constructed from the 100 PB words used during the first day of training in Experiment 1. These words were selected because, in Experiment 1, they were difficult to identify on the first day of training but were reliably identified by the repeated-word subjects on the last day of training. Although all 10 words were always presented in each set of 10 trials, their order within a block of trials was randomly varied.

The pretraining and posttraining test lists each contained 40 novel PB words, plus the 10 words that were used to train subjects in the repeated stimulus condition of the present experiment. These sublists will be referred to as the 40-word subtest and the 10-word subtest, respectively. The words of the 10-word subtest were randomly mixed with those of the 40-word subtest in both the pretraining and posttraining stimulus lists.

Procedure—The same apparatus and general procedures used in Experiment 1 were also used in the present experiment. Subjects were tested in small groups, with a maximum of 6 subjects per session. Subjects were told that they would be listening to single monosyllabic words produced by a text-to-speech system. For the pretraining test, they were instructed to listen carefully to each word and, after each word was presented, to write the word that they heard. Subjects were told to guess if they were uncertain about a word's identity. For the training sessions, subjects were told that after transcribing their response, they would simultaneously be shown the correct response on a CRT monitor, and they would hear a second auditory repetition of the word. All subjects were told that words could be repeated within a list. The posttraining test procedure was identical to that used in the pretraining test (i.e., no feedback was provided to subjects). In all other respects, the training and testing procedures used in the present experiment were identical to those used in Experiment 1.

Results and Discussion

Transcription accuracy was determined according to the procedure used in Experiment 1 for PB words. The principal results concern the performance of the novel- and repeated-stimulus groups during the test sessions. However, because the logic of Experiment 2 requires that the repeated-word list be completely learned prior to the posttraining test, the results for the training session are reported first.

Training data—Performance during training on the novel-and repeated-word lists is summarized in Figure 2. To facilitate comparison with the four blocks of 50 novel words, the 20 repetitions of the 10-word list were grouped into four blocks each containing five repetitions of the 10 words. All of the subjects in the repeated-stimulus condition achieved perfect performance by the third block of trials. For the subjects in the novel-word condition, performance increased significantly from the first training list to the last training list, $F(3, 105) = 52.4$, $MS_e = 0.00273$, $p < .01$. A Newman-Keuls analysis indicated that the percentage of correct word recognition increased from the first training list to the second and

from the third training list to the fourth ($p < .01$). These data demonstrate that subjects in the novel-stimulus condition continued to improve in recognition performance throughout training, whereas those subjects in the repeated-stimulus condition reached perfect performance by the third block of trials.

The superior performance of the subjects trained with repeated words on their training set was apparent in performance on their first 10 trials of training. Repeated-word subjects were able to identify correctly 33% of the first 10 training words they received. These 10 words were presented once, during the pretraining test without feedback, and thus the first set of 10 words in training was already a second encounter with these words for the repeated-word group. By comparison, the first 10 words presented to the novel-word group, at the beginning of training, were never presented to these subjects before. The novel-word subjects were able to identify correctly only 23% of their first 10 novel words, which was significantly worse than was the performance of the repeated-word subjects for the same training block, $F(1, 70) = 8.38$, $MS_e = 0.0215$, $p < .01$. Thus, words produced by the Votrax text-to-speech system were more accurately identified after a single prior presentation in the pretraining test, even though there was no feedback provided for this presentation (cf. Jacoby, 1983).

Testing data—The results for the pre- and posttraining tests are summarized in Table 4. Prior to training, there was a significant difference in performance on the two subtests such that subjects were able to correctly identify only 15% of the 10-word subtest compared to 20% correct responses on the 40-word subtest, $F(1, 70) = 14.55$, $MSe = 0.00644$, $p < .01$. In addition, prior to training there was no performance difference between subject groups and no interaction between subject groups and subtest ($p > .25$).

However, after training, the two subject groups performed quite differently. As expected, performance of the repeated-word subjects was significantly better than was the performance of the novel-word subjects for the 10-word subtest, $F(1, 70) = 310.6$, $MS_e = 0.01716$, $p < .01$. Furthermore, significantly more items were correctly identified in the posttraining 10-word subtest (60%) than in the pretraining subtest (15%), $F(1, 70) = 855.0$, $MS_e = 0.00847$, $p < .01$. Paired comparisons probing the Type of Training × Amount of Improvement interaction indicated that although no significant difference between subject groups was found for the pretraining 10-word subtest, a significant difference between groups was observed for the posttraining 10-word subtest ($p < .01$). These results demonstrate that the subjects in the repeated-word condition learned to recognize words in their 10-word training list much more accurately than did subjects who did not receive those words during training.

Our primary concern is with the amount of generalization learning demonstrated by the two subject groups after training. Performance on the novel 40-word subtest was significantly better after training, $F(1, 70) = 79.9$, $MS_e = 0.0023$, $p < .01$, but there was no significant main effect of type of training on performance, $F(1, 70) = 2.3$, $MS_e = 0.00434$, $p > . 10$. However, the Type of Training × Test Session interaction (pre- vs. posttraining test) was significant, $F(1, 70) = 8.21$, $MS_e = 0.0023$, $p < .01$. A series of paired comparisons probing the interaction revealed that both training conditions produced an improvement in recognition performance ($p < .01$). Unfortunately, the experimental design precludes drawing conclusions about the improvement shown by the repeated-word group: Although their posttraining recognition scores were significantly higher than were their pretraining recognition scores, it is not clear that this result is due to training. A control group that received no training might have demonstrated equal improvement. (The control group in Experiment 1 showed an increase in PB word recognition between the pre- and posttests, presumably as a result of incidental learning in the pretest.) Alternatively, it would not be

surprising if even a small, repeated set of words produces some amount of generalization learning.

However, the goal of the experimental design was to examine the relative improvement demonstrated by repeated-word and novel-word training. Paired comparisons indicated that, although no reliable difference was observed between the two subject groups in the pretraining 40-word subtest, novel-word training produced significantly better generalization performance than did repeated-word training in the posttraining 40-word subtest ($p < .01$). Moreover, a comparison between pretraining and posttraining results demonstrated that novel-word subjects improved significantly more than did repeated-subjects as a result of training ($p < .01$). In short, subjects trained with novel words displayed significantly greater generalization learning than did subjects trained with a repeated set of easily learned words, even though the number of stimuli presented to each group of subjects was equivalent.

Two conclusions can be drawn from this pattern of results. First, novel-word training produced significantly more generalization learning than did repeated-word training. Thus, the generalization performance shown by the novel-stimuli training group was not due to simple exposure to the mechanical sound of synthetic speech. Second, the results of the first experiment indicate that if the repeated set of stimuli represents a relatively large sample of different sentences or words, there is little difference in the effects of training with repeated stimuli compared to training with novel stimuli. However, when the size of the set of repeated items is reduced, as in the second experiment, differences in the effects of novel- and repeated-stimulus training can be observed. Performance on novel items is better if subjects are trained on novel items or on a sufficiently large set of repeated items. This suggests that generalized perceptual learning of synthetic speech may be a consequence of sampling the space of possible stimuli such that the samples represent an adequate description of the structural properties of the speech.

General Discussion

The variability of the acoustic–phonetic structure of synthetic speech generated by the Votrax text-to-speech system is lawful and context-conditioned; moreover, this acoustic–phonetic structure, in principle, is systematically related to the acoustic–phonetic structure of English. Thus, subjects in the present study were faced with the problem of mapping a distorted, but systematic, set of acoustic–phonetic cues onto a previously well-learned set of relations between natural acoustic–phonetic cues and lexical representations in memory. The synthetic speech produced by any text-to-speech system is governed by a set of rules that describe the use of particular phonemes or allophones in specific contexts. But the acoustic–phonetic structure of synthetic speech does not incorporate all the rich and redundant context-conditioned variability that represents natural speech (Pisoni, Nusbaum, & Greene, 1985). Instead, the acoustic–phonetic structure of synthetic speech is constrained much more severely and is limited to a small, fixed inventory of sounds. Thus, in learning to recognize Votrax-generated words and sentences, listeners are really learning to map the limited sound inventory of the Votrax speech onto already well-known phonetic categories, and they are also learning to recognize sequences of these segments as words. Because the inventory of sounds produced by the Votrax is quite limited and the sounds are systematically related to each other through acoustic–phonetic and phonological constraints, listeners may have been able to learn the acoustic–phonetic structure of the synthetic speech from a relatively small set of repeated exemplars in the first experiment.

In interpreting the present results, it is important to note that mere familiarity with the mechanical sound of Votrax was not sufficient to improve intelligibility. Listeners are clearly not simply becoming accustomed to the unusual sound of synthetic speech or to the

sound of Votrax speech, in particular. Rather, listeners are learning specific, detailed acoustic–phonetic information about the structural properties of the speech that is produced by the rule system.

Further support for the conclusion that listeners are learning specific structural properties of the synthetic speech produced by the Votrax system comes from the results of the first experiment comparing perceptual learning for subjects trained on isolated words and subjects trained on fluent sentences. Word-trained and sentence-trained subjects both clearly displayed better recognition of isolated words after training. This indicates that, after training, subjects were more accurate in mapping sound sequences produced by Votrax onto the intended lexical representations. Despite the differences in training materials, both groups of subjects demonstrated similar improvement in word recognition performance. This indicates that sentence-trained subjects did not learn to treat fluent sentences of synthetic speech as holistic entities with a qualitatively different (and more complex) pattern structure than isolated words. However, beyond basic improvements in recognizing words, it appears that the sentence-trained subjects learned something more about perceiving synthetic speech: They learned how to recognize words in sentences, a perceptual skill that was not conferred by training with isolated words alone.

This perceptual skill might be the ability to segment fluent synthetic speech by recognizing words one at a time in the order in which they were produced. According to this segmentation strategy, the beginnings and endings of words are not located by explicit word boundary cues or acoustic information in the speech waveform but simply by the process of serial word recognition. Recognition of the first word in a sentence indicates the beginning of the next word and so on. Another way of expressing this is to say that the sentence-trained subjects learned to recognize words in the absence of word boundary cues that may be expected to be present in natural speech. Conversely, the word-trained subjects, when presented with Votrax-generated sentences, expected the familiar boundary cues, and the absence of these cues impaired word recognition performance for these subjects.

This interpretation suggests that perceptual learning of isolated words might show more effective transfer to recognition of words in fluent speech, if the fluent speech contains explicit segmentation information. For example, with distorted natural speech, such as English produced by a native Japanese speaker, perceptual learning of isolated words might be sufficient for improving recognition of words in fluent sentences, assuming that all languages provide an overlapping set of lexical segmentation cues. Similarly, perceptual learning of isolated words might show greater transfer to recognition of fluent speech when coded natural speech is used, because coded natural speech provides a much more veridical representation of sentence-level phenomena than does synthetic speech generated by rule. In either case, we would expect that learning to recognize isolated lexical patterns should yield better recognition of words in sentences, provided that there is explicit segmentation information present in the speech.

According to many current theories of word perception (e.g., Cole & Jakimik, 1980; Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986), fluent natural speech requires no explicit process for dividing speech into word-size units that are then matched against lexical representations. Instead, segmentation is a direct by-product of the recognition process (e.g., Pisoni, 1978; Reddy, 1976). For the sake of argument, suppose that these theories are correct and that segmentation is a consequence of word recognition and not a necessary antecedent. The sentences produced by the Votrax text-to-speech system are simple concatenations of isolated words, and thus, in these sentences there are no explicit acoustic cues to word boundaries, as have sometimes been observed in natural speech (e.g., Nakatani & Dukes, 1977). Thus, these sentences represent precisely the type of stimuli that

would be expected during word recognition by these theories. Any improvements in recognizing isolated words should directly improve recognition of words in sentences as well. But in spite of their improved recognition of isolated words, the word-trained subjects in our experiment did not, with one exception, show any improvements in recognizing words in fluent sentences, relative to the control subjects. There was just one case in which word-trained subjects were superior to control subjects in recognizing words in fluent sentences: Word-trained subjects displayed better recognition performance for words that followed the determiner “the” in Haskins sentences. Because the sentential position of each word was marked on the response sheets, and the determiners were also printed on these sheets, the subjects were given explicit segmentation information for words following “the.” Thus, when word-trained subjects knew the location of the beginning of a content word in a fluent sentence, they were able to recognize it more accurately after training. This finding demonstrates that the major difficulty experienced by word-trained subjects was in segmenting words from a fluent sentential context. By extension, sentence-trained subjects who clearly learned to recognize words more accurately in sentences really learned to recognize words in fluent speech without explicit segmentation cues. This raises an interesting question. If the general assumption of most theories of auditory word recognition is correct, and segmentation is a consequence of recognition and not its antecedent, then why should subjects need to be trained with fluent speech to recognize words in fluent speech? Similarly, if these theories are correct, why did word-trained subjects only display improved word recognition for sentences when some segmentation information was provided? One suggestion is that this fundamental assumption is wrong: Segmentation is not a consequence of the recognition process but is perhaps an important antecedent or corollary of word recognition.

Taken together, the pattern of results obtained in these experiments suggests that listeners do not normally recognize words one at a time, in the order by which they are produced, as a means of locating word boundaries. Recently, Quene (1985) showed that adding acoustic word boundary cues to synthetic speech does enhance the intelligibility of the synthetic speech. Thus, there may be word segmentation cues in natural speech that might aid in the process of word recognition, and the absence of these cues from Votrax-generated synthetic speech may, in part, impair the ability of listeners to recognize this speech. If segmentation information does indeed play an important role in the recognition process, then most of the current theories of auditory word recognition are based on a fundamentally incorrect assumption and would require considerable revision (Grosjean, 1985; Grosjean & Gee, 1987).

One response to the claim that prosodic and segmentation cues play an important role in word recognition might be to simply incorporate these cues into extant theories of word recognition. For example, word beginnings or endings might be signified by boundary cue detectors that would have a role similar to auditory feature detectors, except that instead of signaling phonetic information to the system (cf. McClelland & Elman, 1986), these boundary detectors would directly fire to the lexical level to indicate the start or end of a word. Although it clearly would not be difficult to add these cues to a theory of word recognition either as part of the lexical representations or as an explicit signaling mechanism, this may not be the appropriate way to incorporate this information. The addition of these cues might allow a theory to emulate human performance, but it would not be dictated on computational grounds. Including these cues would not enhance the performance of the model, except in terms of emulating humans. Instead, it seems that the need for segmentation cues should dictate a different approach to word recognition other than the current, strictly linear, word-byword strategy (e.g., see Grosjean & Gee, 1987).

Finally, it is not clear whether repetition and novel-stimulus training in Experiment 1 would have produced equivalent effects if subjects in the repetition training condition had achieved asymptotic performance during the training sessions. If subjects extract all the information from a fixed set of stimuli so that no more overt learning occurs, and repetition training and novel-stimulus training is continued on from that point, will the two types of training be equivalent? Once learning reaches asymptotic levels in a repetition condition, the effects of repetition learning on a generalization test may level off, whereas novel-stimulus training may continue to produce increasingly better performance on a generalization test. This is an issue that warrants investigation in future research on perceptual learning.

The results of the present experiments demonstrate the importance of studying generalization learning for stimuli that are lawfully related to previously well-learned stimulus structures and that are internally coherent and involve context-conditioned variability. Moreover, these results indicate that it is not always advisable to infer similarities between the processes of visual pattern recognition and speech perception. The processes that mediate perceptual learning appear to be linked directly to the type of pattern structures that are presented in different modalities and, as a consequence, perceptual learning may take different forms for different types of stimulus sets across sensory systems. It is important to begin to characterize what these differences and similarities are and how they may affect the processes of perceptual learning in order to develop more general theories of perceptual learning. Simple exemplar-based models (e.g., Jacoby, 1983) or stimulus-response associating models (e.g., Shiffrin & Schneider, 1977) may be inadequate for the task of representing the full range of complexity presented by perceptual learning in different modalities for different types of stimuli. In future work it will be necessary to investigate systems that are capable of representing the rich structural relations that exist among different tokens of stimuli and that represent perceptual learning as a more general process of complex skill acquisition (e.g., Grossberg, 1982; Kolers & Roediger, 1984).

Acknowledgments

This research was supported in part by Air Force contract AF-F-33615-83-K-0501 with the Air Force Systems Command, Air Force Office of Scientific Research, through the Aerospace Medical Research Laboratory, Wright-Patterson Air Force Base, Ohio; in part by NIH Grant NS-12719; and in part by NIH Training Grant NS-07134-06 to Indiana University in Bloomington.

We thank Kimberley Baals and Michael Stokes for their valuable assistance in collecting and scoring the data; Robert Bernacki, Michael Dedina, and Jerry Forshee for technical assistance.

References

Cole, RA.; Jakimik, J. A model of speech perception. In: Cole, RA., editor. Perception and production of fluent speech. Erlbaum; Hillsdale, NJ: 1980. p. 133-164.

Dukes WF, Bevan W. Stimulus variation and repetition in the acquisition of naming responses. Journal of Experimental Psychology. 1967; 74:178–181. [PubMed: 6048453]

Egan JP. Articulation testing methods. Laryngoscope. 1948; 58:955–991. [PubMed: 18887435]

Elman, J.; McClelland, J. Exploiting lawful variability in the speech wave. In: Perkell, JS.; Klatt, DH., editors. Invariance and variability in speech processes. Erlbaum; Hillsdale, NJ: 1986. p. 360-380.

Grosjean F. The recognition of words after their acoustic offset: Evidence and implications. Perception & Psychophysics. 1985; 38:299–310. [PubMed: 3831907]

Grosjean F, Gee JP. Prosodic structure and spoken word recognition. Cognition. 1987; 25:135–155. [PubMed: 3581724]

Grossberg, S. Studies of mind and brain: Neural principles of learning, perception, development, cognition, and motor control. Reidel Publishing; Boston: 1982.

House AS, Williams CE, Hecker MHL, Kryter K. Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America. 1965; 37:158–166. [PubMed: 14265103]

Institute of Electrical and Electronics Engineers (IEEE). IEEE recommended practice for speech quality measurements (IEEE Report No. 297). Author; New York: 1969.

Jacoby LL. Perceptual enhancement: Persistent effects of an experience. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1983; 9:21–38.

Kolers PA, Magee L. Specificity of pattern-analyzing skills in reading. Canadian Journal of Psychology. 1978; 32:43–51.

Kolers PA, Roediger HL III. Procedures of mind. Journal of Learning and Verbal Behavior. 1984; 23:425–449.

Liberman AM. The grammars of speech and language. Cognitive Psychology. 1970; 1:301–323.

Liberman AM, Cooper FS, Shankweiler D, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 84:452–471.

Markel, JD.; Gray, AH, Jr.. Linear prediction of speech. Springer-Verlag; New York: 1976.

Marslen-Wilson WD, Welsh A. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology. 1978; 10:29–63.

Massaro, DW.; Oden, GC. Speech perception: A framework for research and theory. In: Lass, NJ., editor. Speech and language: Advances in basic research and practice. Vol. Vol. 3. Academic Press; New York: 1980.

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18:1–86. [PubMed: 3753912]

Nagata H. Quantitative and qualitative analysis of experience in acquisition of a miniature artificial language. Japanese Psychological Research. 1976; 18:174–182.

Nakatani LH, Dukes KD. Locus of segmental cues for word juncture. Journal of the Acoustical Society of America. 1977; 62:714–719. [PubMed: 903512]

Nusbaum HC, Pisoni DB. Some constraints on the perception of synthetic speech. Behavior Research Methods, Instruments, & Computers. 1985; 17:235–242.

Nye, PW.; Gaitenby, J. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories; New Haven, CT: 1974. p. 169-190.report on speech research, SR-37/38

Palermo DS, Parish M. Rule acquisition as a function of number and frequency of exemplars presented. Journal of Verbal Learning and Verbal Behavior. 1971; 10:44–51.

Pisoni, DB. Speech perception. In: Estes, WK., editor. Handbook oflearning and cognitive processes: Vol. 6, Linguistic in cognitive theory. Erlbaum; Hillsdale, NJ: 1978. p. 167-233.

Pisoni DB. Speech perception: Some new directions in research and theory. Journal of the Acoustical Society of America. 1985; 78:381–388. [PubMed: 4031245]

Pisoni DB, Nusbaum HC, Greene BG. Perception of synthetic speech generated by rule. Proceedings of the IEEE. 1985; 73:1665–1676.

Posner MI, Keele SW. On the genesis of abstract ideas. Journal of Experimental Psychology. 1968; 77:353–363. [PubMed: 5665566]

Quene, H. Progress Report of the Institute of Phonetics. University of Utrecht; The Netherlands: 1985. Consonant duration as a perceptual boundary cue in Dutch; p. 15-34.

Reddy DR. Speech recognition by machine: A review. Proceedings of the IEEE. 1976; 64:501–531.

Schwab EC, Nusbaum HC, Pisoni DB. Some effects of training on the perception of synthetic speech. Human Factors. 1985; 27:395–408. [PubMed: 2936671]

Shiffrin RM, Schneider W. Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review. 1977; 84:127–190.

Stevens KN, Blumstein SE. Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America. 1978; 64:1358–1368. [PubMed: 744836]

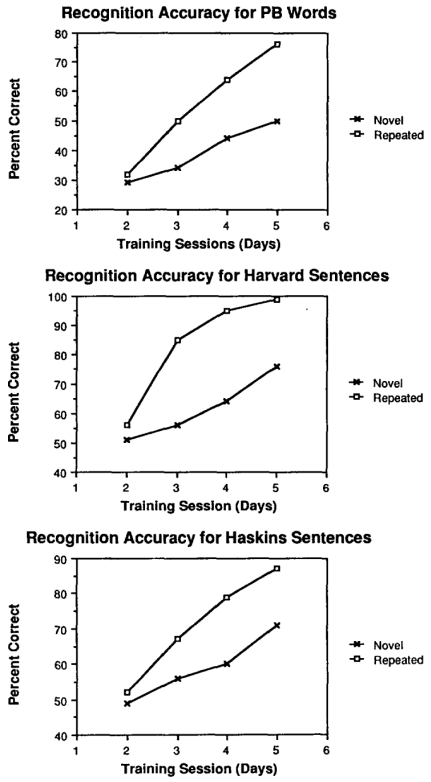

Figure 1.
Transcription accuracy for words in the phonetically balanced word lists (top panel), Harvard sentences (middle panel), and Haskins sentences (bottom panel) for each day of training in Experiment 1. (*Note*. Days 1 and 6 were pre- and posttraining test sessions.)

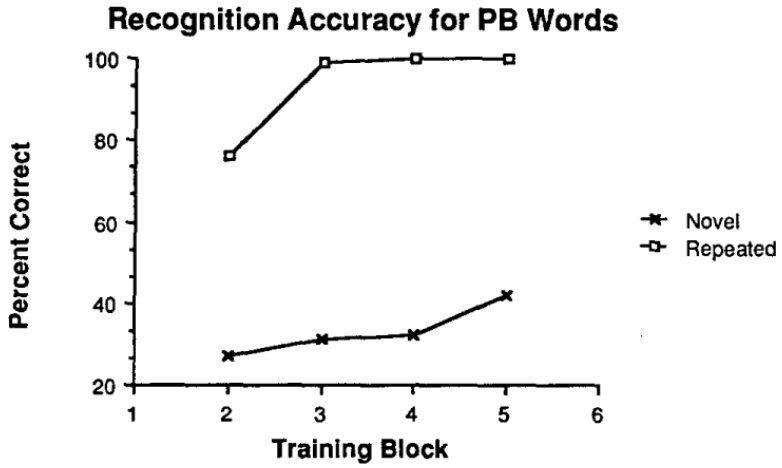

Figure 2.
Transcription accuracy for words in the phonetically balanced word lists for each training list of Experiment 2. (*Note*. Lists 1 and 6 were pre- and posttraining test lists.)

Table 1

Transcription Accuracy for Words in the Pretraining and Posttraining Isolated-Word Tasks (Percentage Correct)

Condition	Pretraining	Posttraining	Change (% points)
Modified rhyme test			
Control	65.5	66.6	1.1
Novel words	64.6	75.8	11.2
Repeated words	65.1	75.6	10.5
Novel sentences	67.4	76.4	9.0
Repeated sentences	64.9	71.3	6.4
Phonetically balanced words			
Control	26.8	36.0	9.2
Novel words	25.5	47.3	21.8
Repeated words	24.2	48.2	24.0
Novel sentences	25.6	47.4	21.8
Repeated sentences	25.8	48.1	22.3

Table 2

Transcription Accuracy for Words in the Pretraining and Posttraining Sentence Tasks (Percentage Correct)

Condition	Pretraining	Posttraining	Change (% points)
Harvard sentences			
Control	49.2	38.0	−11.2
Novel words	44.7	34.7	−10.0
Repeated words	40.6	40.6	0
Novel sentences	45.3	62.3	17.0
Repeated sentences	44.4	58.4	14.0
Haskins sentences			
Control	24.0	42.3	18.3
Novel words	19.0	41.7	22.7
Repeated words	21.3	41.7	20.4
Novel sentences	25.6	65.0	39.4
Repeated sentences	25.9	69.3	43.4

Table 3

Transcription Accuracy for Words in the Pretraining and Posttraining Haskins Sentences as a Function of Word Position (Percentage Correct)

	Test session					
	Pretraining		Posttraining		Improvement	
Condition	**After "the"**	**After open-class**	**After "the"**	**After open-class**	**After "the"**	**After open-class**
Control	30.8	17.1	56.7	27.9	25.9	10.8
Novel words	25.0	12.9	62.9	20.4	37.9	7.5
Repeated words	27.7	15.0	60.8	22.7	33.1	7.7
Novel sentences	32.9	18.3	73.3	56.7	40.4	38.4
Repeated sentences	36.4	15.5	81.4	57.3	45.0	41.8

Note. The category *After "the"* contains words that were presented immediately after the word "the." The category *After open-class* contains words that were presented immediately after an open-class word (see text for further details).

Table 4

Transcription Accuracy for Words in the Pretraining and Posttraining Tests (Percentage Correct)

	Test session			
	Pretraining		Posttraining	
Condition	10-word subtest	40-word subtest	10-word subtest	40-word subtest
Repeated	16.7	20.7	97.2	25.6
Novel	13.9	20.1	23.1	29.5