REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release; Distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 89-0963

6a. NAME OF PERFORMING ORGANIZATION
The University of Chicago

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Directorate of Life Sciences
Air Force Office of Scientific Research

6c. ADDRESS *(City, State and ZIP Code)*
970 East 58th Street
Chicago, Illinois 60637

7b. ADDRESS *(City, State and ZIP Code)*
Building 410
Bolling AFB, Washington, D.C. 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL *(If applicable)*
NL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-87-0272

8c. ADDRESS *(City, State and ZIP Code)*
Building 410
Bolling AFB
Washington, D.C. 20332-6448

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.
61102F	2313	A4	

11. TITLE *(Include Security Classification)* (unclassified)
Attention and vigilance in speech perception

12. PERSONAL AUTHOR(S)
Howard C. Nusbaum

13a. TYPE OF REPORT
Final Technical Report

13b. TIME COVERED
FROM 7/1/87 TO 12/31/88

14. DATE OF REPORT *(Yr., Mo., Day)*
89,06,14

15. PAGE COUNT
72

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.
05	10	09

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Attention, speech perception, syllables, phonemes, talker normalization, perceptual learning, synthetic speech, cognitive load

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This report describes research carried out in three related projects investigating the function and limitations of attention in speech perception. The projects were directed at investigating the distribution of attention in time during phoneme recognition, perceptual normalization of talker differences, and perceptual learning of synthetic speech. The firs project demonstrates that in recognizing phonemes listeners attend to earlier and later phonetic context, even when that context is in another syllable. The second project demonstrated that there are two mechanisms underlying the ability of listeners to recognize speech across talkers. The first, structural estimation, is based on computing a talker-independent representation of each utterance on its own; the second, contextual tuning, is based on learning the vocal characteristics of the talker. Structural estimation requires more attention and effort than contextual tuning. The final project examined the attentional demands of synthetic speech and how they change with perceptual learning. The results demonstrated that the locus of attentional demands in perception of synthetic speech is in

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☒ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr. Alfred R. Fregly

22b. TELEPHONE NUMBER *(Include Area Code)*
(202) 767 5021

22c. OFFICE SYMBOL
AFOSR/NL

Block 19 Continued:

recognition rather than storage or recall of synthetic speech. Moreover, perceptual learning increases the efficiency with which listeners can use spare capacity in recognizing synthetic speech and this effect is not just due to increased intelligibility. Our results suggest that perceptual learning allows listeners to focus on the relevant acoustic-phonetic properties of a particular, synthetic talker.

AFOSR-TR- 89-0963

ATTENTION AND VIGILANCE IN SPEECH PERCEPTION

Howard C. Nusbaum
Speech Research Laboratory
Department of Psychology
The University of Chicago
5848 South University Avenue
Chicago, Illinois 60637

23 June 1989

Final Report for Period 1 July 1987 — 31 December 1988

Distribution Statement

Prepared for

DIRECTORATE OF LIFE SCIENCES
Air Force Office of Scientific Research
Bolling AFB
Washington, D. C. 20332-6448

Speech Research Laboratory Personnel

Howard C. Nusbaum, Ph.D.	Assistant Professor and Director
Jenny DeGroot, B.A.	Graduate Research Assistant
Lisa Lee, B.A.	Graduate Research Assistant
Todd M. Morin, B.A.	Graduate Research Assistant

Summary

This report describes the research that we have carried out to investigate the role of attention in speech perception. In order to conduct this research, we have developed a computer-based perceptual testing laboratory in which an IBM-PC/AT controls experiments and presents stimuli to subjects, and individual Macintosh Plus subject stations present instructions to subjects and collect responses and response times. Using these facilities, we have completed a series experiments in three projects. These experiments examine the integrality of syllables and syllable onsets in speech (Project 1), the attentional demands incurred by normalization of talker differences in vowel perception (Project 2), and the effects on attention of perceptual learning of synthetic speech (Project 3). The results of our first project demonstrate that adjacent phonemes are treated as part of a single perceptual unit, even when those phonemes are in different syllables. This suggests that, although listeners may attend to a phonemic level of perceptual organization, syllable structure and syllable onsets are less important in recognizing consonants than is the acoustic-phonetic structure of speech. This finding argues against several recent claims regarding the importance of syllable structure in the early perceptual processing and recognition of speech.

Our second project provides evidence for the operation of two different mechanisms mediating the normalization of talker differences in speech perception. When listeners hear a sequence of vowels, syllables, or words produced by a single talker, recognition of a target phoneme or word is faster and more accurate than when the stimuli are produced by a mix of different talkers. This demonstrates the importance of learning the vocal characteristics of a single talker for phoneme and word recognition (i.e., contextual tuning). However, even though there are reliable performance differences in speech perception between the single- and multiple-talker conditions, these differences are small, suggesting the operation of a mechanism that can perform talker normalization based on a single token of speech (i.e., structural estimation). Recognition based on this mechanism is slower and less accurate than is recognition based on contextual tuning. Furthermore, contrary to recent claims, there is no performance advantage in recognizing vowels in CVC context compared to isolated vowels and consonant context does not facilitate perceptual normalization. Finally, we found that the operation of the structural estimation mechanism places demands on the capacity of working memory which are not imposed by contextual tuning.

In our third project, we investigated the effects of perceptual learning of synthetic speech on the capacity demands imposed by synthetic speech during serial-ordered recall and speeded word recognition. Moderate amounts of training on synthetic speech produces significant improvements in recall of words generated by a speech synthesizer. In addition, increasing memory load by

visually presenting digits prior to the spoken words decreased the amount of synthetic speech recalled. However, there was no interaction between memory preload and training indicating that the representation of synthetic speech does not require any more or less capacity after training. The pattern of results is much the same for a speeded word recognition task carried out before and after training with one significant exception: There is a significant interaction between cognitive load and training such that training allows listeners to use surplus cognitive load more effectively. Our findings suggest that if training changes the attentional demands of perceiving synthetic speech, these changes occur at the level of perceptual encoding rather than in the storage of words. Moreover, it appears that the effects of training are directly on the use of capacity rather than indirectly through changes in intelligibility. A comparison of the effects of manipulating cognitive load on speeded word recognition in high- and low-intelligibility synthetic speech does not yield a similar interaction.

Taken together, our research has begun to specify some of the functions and the operation of attention in speech perception. A number of new experiments are suggested by our current and anticipated results. These experiments will provide basic information about the cue information used in normalization of talker differences, the limits of integrality among phonemes and within other units, changes in attentional limitations imposed by recognition of synthetic speech following training, and habituation and vigilance effects in speech perception.

Conference Presentations and Publications

Nusbaum, H. C. Understanding speech perception from the perspective of cognitive psychology. To appear in P. A. Luce & J. R. Sawusch, (Eds.), *Workshop on spoken language*. In preparation.

Nusbaum, H. C. (1988). *Attention and effort in speech perception*. Air Force Workshop on Attention and Perception, Colorado Springs, CO, September.

Nusbaum, H. C., & Morin, T. M. (1988). *Perceptual normalization of talker differences*. Psychonomics Society, Chicago, IL, November.

Nusbaum, H. C., & Morin, T. M. (1988). *Speech perception research controlled by microcomputers*. Society for Computers in Psychology, Chicago, IL, November.

DeGroot, J., & Nusbaum, H. C. (1989). *Syllable structure and units of analysis in speech perception*. Acoustical Society of America, Syracuse, May.

Lee, L., & Nusbaum, H. C. (1989). *The effects of perceptual learning on capacity demands for recognizing synthetic speech*. Acoustical Society of America, Syracuse, May.

Nusbaum, H. C., & Morin, T. M. (1989). *Perceptual normalization of talker differences*. Acoustical Society of America, Syracuse, May.

Attention and Vigilance in Speech Perception
Final Report: 7/87-12/88

I. Introduction

In listening to spoken language, subjectively we seem to recognize words with little or no apparent effort. However, over twenty years of research has demonstrated that speech perception does not occur without attentional limitations (see Moray, 1969; Nusbaum & Schwab, 1986; Treisman, 1969). Given that there are indeed attentional limitations on the perceptual processing of speech, what is the nature of these limitations and why do they occur?

We have begun to examine more carefully the role of attention in speech perception and how attentional limitations can be used to investigate the processes that mediate the recognition of spoken language. To date, we have investigated three specific questions: (1) What perceptual units are used by the listener to organize and recognize speech? (2) How do listeners accommodate variability in the acoustic representations of different talkers' speech? (3) What are the effects of perceptual learning on the capacity demands incurred by the perception of synthetic speech?

These three specific questions represent starting points for investigating three very broad issues that are fundamental to understanding the perceptual processing of speech. How does the listener represent spoken language? How does the listener map the acoustic structure of speech onto these mental representations? And finally, what is the role of learning in modifying the recognition and comprehension of spoken language? The first two questions are important because of the lack of acoustic-phonetic invariance in speech. If acoustic cues mapped uniquely and directly onto linguistic units, we would have little difficulty understanding the mechanisms that mediate speech perception. But the many-to-many relationship between the acoustic structure of speech and the linguistic units we perceive has not been explained completely by any theoretical accounts to date. In order to understand how the human listener perceives speech, we must understand the types of units used to organize and recognize speech and we must understand the recognition processes that overcome the lack of acoustic-phonetic invariance.

The third question regarding the perceptual learning of speech has received less attention in general speech research. While numerous studies have investigated the development of speech perception in infants and young children (see Aslin, Pisoni, & Jusczyk, 1983), there is much less known about the operation of perceptual learning of speech in adults, in which there is a fully developed language system. Based on subjective experience, it seems that adult listeners are much less capable than infants of modifying their speech production system to learn a new language. However, adult listeners can acquire new phonetic contrasts not present in their native language (Pisoni, Aslin, Perey, & Hennessy, 1982). Furthermore, listeners can learn to recognize synthetic speech, despite its impoverished acoustic-phonetic structure (Greenspan, Nusbaum, & Pisoni, in press; Schwab, Nusbaum, & Pisoni, 1985). By understanding how the listener's perceptual system changes as a function of training, we will learn a great deal more about the processes that mediate speech perception.

II. Instrumentation Development

In order to carry out our research on the role of attention in speech perception, it was necessary to develop an on-line, real-time perceptual testing laboratory. Because this development effort has required a substantial amount of time, and is critical .o the implementation and successful completion of our research program, we will outline our development efforts briefly. In the past, speech research has been conducted under the control of PDP-11 laboratory minicomputers. However, the cost of these systems and their computational limitations on CPU speed, memory size, and I/O bandwidth have made them unattractive for controlling more complex experimental paradigms by comparison with the more modern MicroVax. Unfortunately, the cost of this system has been too great for a newly developing laboratory.

Our research program depends on the ability to present speech signals to listeners and collect response times with millisecond accuracy from subjects. The basic system that we have developed consists of an experiment-control computer that is connected to individual subject stations. We chose the IBM-PC/AT as our experiment control system because it provided a cost-effective system that is capable of digitizing and playing speech from disk files. The subject stations are Macintosh Plus computers which are capable of maintaining a millisecond timer and collecting keyboard responses with millisecond accuracy. Also, this system has a vertical retrace interrupt which allows us to start timing a response interval from the presentation of a visual stimulus.

The software we have developed for the experiment control system and subject stations distributes the demands of an experiment among the different microcomputers so that no single system must bear the entire computational load. The PC/AT sequences and presents stimuli to subjects and it sends a digital signal to the subjects stations to start a timer or to present a visual display. This signal is presented by a digital output line to the mouse port of the Macintosh Plus which the Macintosh can detect with minimal latency. Thus, in a trial, the AT will send a signal to start timing a response and then it will play out a speech signal. Each of the Macintosh computers starts a clock and then waits for a subject's keypress. The keypress and response time are then sent back to the AT over a serial line for storage in a disk file. We have calibrated our subject station timers against the PC/AT and we have found them accurate to the millisecond, More recently, we have replicated an experiment with stimuli that were used with an older PDP-11 computer and the results from the two experiments were within milliseconds of each other.

In spite of the success of our instrumentation development, the limitations of using an IBM-PC/AT have become clear. The number of stimuli that can be used in an experiment is limited by the driver software for the D/A system. Only relatively short dichotic stimuli can be played from disk and the memory limitations of the segmented architecture of the AT limits the size of stimuli held in memory. Thus, while this system is adequate for experiments involving small numbers of stimuli or relatively short stimuli, for more complex experiments involving dichotic presentations of long word or sentence-length materials or large stimulus sets, it will be necessary to move to a MicroVax or Macintosh II for experiment control. Since we designed the system to be modular and the software is all written in C and is thus transportable directly to other computers, moving to a more powerful computer and operating system will only require minor changes in the existing experiment control software and no changes in the subject stations.

III. Project 1: Perceptual Integrality of Perceptual Units in Speech

What is the basic unit of perception used by listeners in recognizing speech? This is an important question because in order to understand speech perception we must know *what* listeners recognize, as well as *how* recognition takes place. Although we typically hear speech as a sequence of words, we must have some type of segmental or sublexical representation, since we are able to recognize and reproduce or transcribe nonwords, and because we can always learn new words that have never been heard before (Pisoni, 1981). Candidates for the unit of perceptual analysis have been numerous including: acoustic properties, phonetic features, the context-conditioned allophones, phonetic segments, phonemes, and syllables (see Pisoni, 1978). However, the strongest linguistic arguments have been made in favor of both the phoneme (Pisoni, 1981) and the syllable or subsyllabic structure (Fudge, 1969; Halle & Vergnaud, 1980).

The syllable structure view posits that syllables are composed of onsets and rimes. The onset consists of all the consonants before vowel in a syllable or the onset can be null. The rime consists of the vowel (called the peak or nucleus) followed by the coda or offset which consists of all the consonants (if any) following the peak. Treiman (1983) has argued for the psychological reality of this type of syllabic organization based on the ability of children to play word games like pig latin that require the segmentation of words into different pieces. Onset-rime divisions are easier to make than divisions within onsets.

More recently Treiman, Salasoo, Slowiaczek, & Pisoni (1982) used a phoneme monitoring task to demonstrate that listeners were slower to recognize phoneme targets when they occurred within consonant clusters as onsets, than when the phoneme targets occurred as the only segment in the onset. Similarly, Cutler, Butterfield, and Williams (1987) also claimed to find support for the perceptual reality of onset structures in recognition of speech. However, performance in both of these experiments was quite poor: Accuracy in the experiments described by Cutler et al. was around 80% correct. In the Treiman et al. (1982) study, response times to recognize fricative targets were in the range of 900 to 1000 msec which are much longer RTs than the 300-500 msec RTs typically found in phoneme monitoring studies. Because of these performance problems, it is simply not clear what subjects were doing in these experiments and the results may reflect more the operation of metalinguistic awareness of language structure than the operation of normal perceptual coding and recognition processes. Nonetheless both sets of studies provided some evidence supporting the hypothesis that syllabic onsets form an integral perceptual unit.

Experiment 1.1: Stop Consonant Identification in Fricative Contexts

The purpose of our first experiment was to test the claim that syllable onsets are perceptual units that are integral in speech recognition. The methodology used in the Treiman et al. and Cutler et al. studies was based on the assumption that subjects should be slower to recognize a single phoneme in a complex onset (e.g., /s/ in /st/) than when the phoneme is presented alone as the onset. One problem with this approach is that the differences in response times observed in these studies could have been due to acoustic-phonetic differences in the stimuli. For example, in the Treiman et al. study, listeners heard CV, CVC, and CCV stimuli and responded yes or no based on the presence or absence of a target fricative. However, the response time and accuracy differences could reflect differences in the intelligibility of the stimuli among these syllable types rather than reflecting differences in the recognition of segments in onsets.

The present study was designed to use a different methodology for testing the claim that syllable onsets form an integral perceptual unit. According to Garner (1974), if two dimensions of a perceptual unit are integral, and subjects are asked to make judgments about one of the dimensions, variation in the other dimension should affect response times. If variation in a second dimension is correlated with variation in the target dimension (the correlated condition), subjects should be faster to judge the target dimension than if the second dimension is held constant (the unidimensional condition). Also, irrelevant (uncorrelated) variation in the second dimension should slow responses to the target dimension (the orthogonal condition). On the other hand, if the two dimensions are separable in perception of the unit, variation in a second dimension could be filtered out by the subject and ignored. Thus, with separable dimensions, there should be no difference between response times in orthogonal and unidimensional conditions. Response time for the correlated condition could be the same as the response time to the unidimensional condition, or it could be faster due to a redundancy gain.

Wood and Day (1975) demonstrated that listeners treat the consonant and vowel in a CV syllable as two dimensions of a perceptually integral unit. The speed of judgments of the identity of the consonant were affected by manipulations of the identity of the vowel. In the present experiment, we investigated the perceptual integrality of syllable onsets and syllables. The two "dimensions" we manipulated are the identity of a stop consonant (i.e., /p/ or /t/) and the identity of a preceding fricative (i.e., /s/ or /ʃ/) in syllables such as spa, sta, shpa, shta. For these syllables, subjects judged the identity of the stop consonant in unidimensional, correlated, and orthogonal conditions. If the onset is perceptually integral, subjects should respond faster in the correlated condition than in the unidimensional condition and they should respond more slowly in the orthogonal condition than in the unidimensional condition. On the other hand, if the onset is separable and not a single perceptual unit, there should be no difference in response times across these conditions. The advantages to this paradigm over the previous studies are that each stimulus serves as its own control across conditions and that this paradigm is designed specifically to assess the integrality of perceptual dimensions.

Of course, response time differences across these conditions could be due to some type of integrality due to phonetic adjacency, rather than anything specific to the integrality of the syllable onset. Therefore, we included a set of bisyllabic stimuli /isˈp^{h}a/, /isˈt^{h}a/, /iʃˈp^{h}a/, and /iʃˈt^{h}a/ (/i/ is pronounced "ee" and the ' mark means that the syllable following the mark is stressed). These stimuli are important because they contain the exact same fricative-stop sequence as the monosyllabic stimuli. However, for these bisyllabic utterances, the fricative and stop consonant are in different syllables. The fricative is the coda of the first syllable and the stop is the onset of the second syllable. The syllables were produced by stressing the second syllable and aspirating the stop consonant, so that native English listeners would perceive the fricative and stop as segments in different syllables. If syllable onsets are integral perceptual units, the response time differences found for the monosyllabic stimuli should not be observed with these bisyllabic stimuli. Moreover, this experiment tests whether or not an entire syllable (in addition to just the onset) is perceptually integral, since the difference in onset structure is identical to the difference in syllable structure (monosyllabic vs. bisyllabic). If the results indicate that response times to the monosyllabic stimuli display a pattern consistent with integrality while the bisyllabic stimuli display a pattern consistent with separability, we would be unable to determine whether the entire syllable or just the syllable onset was integral, from this experiment alone. However, these results would be consistent with the onset integrality hypothesis as well.

Method

Subjects. The subjects were 18 University of Chicago students and residents of Hyde Park, aged 18-28. All the subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hou~ for their participation.

Stimuli. The stimuli were 8 utterances spoken by a single male talker. Four of these utterances were monosyllables beginning with a fricative-stop consonant cluster: /spa/, /sta/, /ʃpa/, and /ʃta/. The other four items — /isˈp^{h}a/, /isˈt^{h}a/, /iʃˈp^{h}a/, and /iʃˈt^{h}a/ — contained the same fricative-stop sequences, but with the two consonants in different syllables. The bisyllabic words were stressed on the second syllable, and the stop was aspirated. In English, only syllable-initial stops are aspirated; thus, the fricative and stop in /isˈp^{h}a/, e.g., are not heard by native English speakers as a syllable-initial consonant cluster.

For the purposes of recording, the test utterances were produced in sequences of similar utterances, for example, "sa, spa, sa". For each test stimulus, several such triads were recorded on cassette tape in a sound-shielded booth. The utterances were digitized at 10 kHz with 12-bit resolution and were low-pass filtered at 4.6 kHz. The stimuli were initially stored as a single digitized waveform on the hard disk of an IBM-PC/AT.

Because natural speech was used, there was some variation in duration and intonation of the utterances. For each test stimulus, a single token was selected from among the several tokens of each of the four monosyllabic and bisyllabic utterances. The selection was based on prosodic similarity as judged by a trained phonetician. The selected tokens were edited with a digital waveform editor with 100 microsec accuracy. Each token was visually inspected and excised into individual waveform files by deleting all acoustic information before the onset of the initial aperiodic noise (for /s/ or /ʃ/) or periodicity (for /i/), and after the end of periodicity (for /ɑ/). After editing, the waveforms were played to ensure that the onset and offset of each nonsense word were not abrupt.

The stimuli were played to subjects over Sennheiser HD-430 headphones at about 76 dB SPL as measured with a single calibration token /spɑ/. Digitized stimuli were converted into speech in real-time under computer control. Each waveform was played at 10 kHz through a 12-bit D/A converter and low-pass filtered at 4.6 kHz.

Procedure. Small groups of one to three subjects were tested in a single experimental session lasting about an hour. Each subject sat in a sound-attenuated booth, facing a Macintosh Plus microcomputer. For 11 of the subjects, the Z key on the Macintosh keyboard (the bottom leftmost character key) was labeled as the **p** response button, and the / key (at the opposite end of the same row of keys) was labeled **t**. For the other 7 subjects, the position of the **p** and **t** labels was reversed.

The subjects were told that on each trial they would hear one token of the specified stimulus set over headphones. They were instructed to determine whether each stimulus contained a /p/ or a /t/ sound, and to press the corresponding key as quickly as possible without sacrificing accuracy. Responses and response times for each subject on each trial were recorded by the Macintosh computer and stored in a file on the IBM-PC/AT.

Subjects participated in a practice block of trials, and three experimental conditions: a correlated-dimensions condition, an orthogonal-dimensions condition, and a unidimensional condition (Garner, 1974). All subjects first received practice with five repetitions of each of the four monosyllables presented in random order. For each practice trial, the choices **p** and **t** appeared on opposite sides of the Macintosh screen, above the corresponding keys. An utterance was presented binaurally, and each subject pressed a response key. After all subjects responded, feedback was presented: An orthographic transcription of the utterance was displayed in the center of the screen (spelled **spa**, **sta**, **shpa**, or **shta**), while the stimulus waveform was presented again over the headphones.

After the practice block, the three experimental conditions were presented in five blocks of trials; each block consisted of 20 repetitions of each stimulus appropriate to that block, presented in random order. No feedback was presented during the experimental blocks and subjects responded using the same response keys and labels as used in the practice block.

Two of the blocks of trials made up the correlated condition. In these blocks, variation in the stop consonant was correlated with variation in the fricative: one stop consonant (e.g., /p/) always occurred with the same fricative (e.g., /s/), and the other stop always occurred with the other fricative. The first correlated block thus consisted of 20 repetitions each of /spɐ/ and /ʃtɐ/, and the second was composed of /ʃpɐ/ and /stɐ/. In the two blocks of unidimensional condition trials, only the stop consonant was varied. The first block consisted of 20 repetitions each of /spɐ/ and /stɐ/, and the second consisted of /ʃpɐ/ and /ʃtɐ/. Finally, a single block of trials was presented in the orthogonal condition. In this condition, both the fricative and stop both varied and 20 repetitions of each of the four monosyllables were presented. The order of conditions was varied across subjects.

After the monosyllables were presented, the equivalent set of unidimensional, correlated, and orthogonal conditions were presented using bisyllabic stimuli. The correlated condition consisted of an /isˈpʰɐ/-/iʃˈtʰɐ/ block and an /iʃˈpʰɐ/-/isˈtʰɐ/ block. In the unidimensional condition blocks, the fricative was constant within a block and the stop consonant was varied. In the orthogonal block, all four bisyllabic stimuli were presented. Each subject received the bisyllabic stimulus conditions in the same order as the monosyllabic, beginning with five practice repetitions of each item, in random order. Again, each experimental block consisted of twenty repetitions of the stimuli for that block, presented in random order and the order of the blocks was varied across subjects.

Results and Discussion

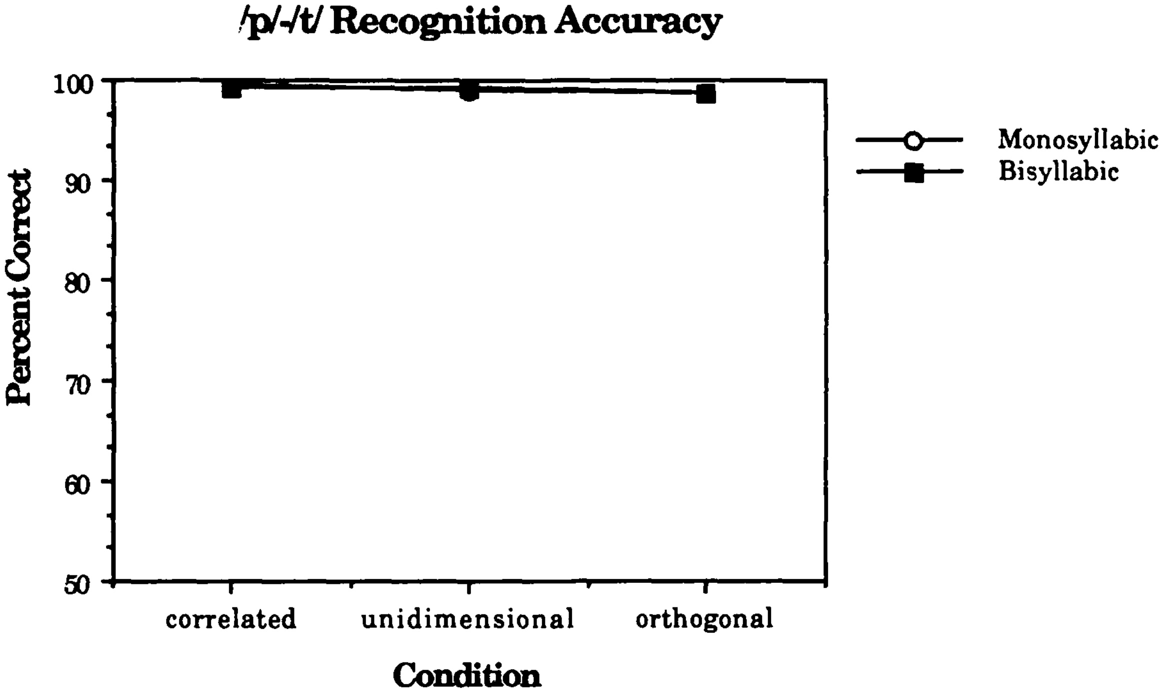

Figure 1.1. Recognition accuracy for stop consonants /p/ and /t/ in unidimensional correlated, and orthogonal conditions when irrelevant contextual variation is in the same syllable (open circles) or a different syllable (closed squares).

Figure 1.1 shows that the mean accuracy in judging the identity of the stop consonant was excellent, about 99% correct for all conditions. There were no statistically significant differences in accuracy among any of the conditions or individual stimuli.

Figure 1.2 shows the mean response times for the /p/-/t/ judgments for monosyllabic and bisyllabic stimuli in unidimensional, correlated, and orthogonal conditions. Response times were affected significantly by condition (unidimensional vs. correlated vs. orthogonal), $F(2,34) = 11.293$, $p < .001$, although there was no effect of syllable structure (monosyllabic vs. bisyllabic), $F(1,17) = .023$, n.s., and the interaction was not significant, $F(2,34) = .081$, n.s. Post-hoc Newman-Keuls analyses showed that response times were fastest in the correlated condition, significantly slower in the unidimensional condition, and slowest in the orthogonal condition ($p < .05$).

Our results indicate that stop consonants and their preceding fricatives are perceived as integral perceptual units, according to Garner's (1974) criteria, regardless of syllable structure. Even though the phonemes are linguistic units by themselves, listeners are unable to identify the stop consonants in these stimuli without processing the fricatives. This finding is consistent with the results reported by Wood and Day (1975) that an adjacent consonant and vowel are perceived as integral, but our overall pattern of results argues against the conclusion that the syllable is the relevant integral unit of analysis. Our results do not show any indication of any difference in the integrality of stops and fricatives as a function of syllable structure or onset structure. If syllables or syllable onsets are perceptually important for recognizing the linguistic structure of speech (e.g., Cutler et al., 1987; Treiman et al., 1982), the stops and fricatives should have been integral in the monosyllabic stimuli, but separable in the bisyllabic stimuli. If the syllable or syllable onset is the primary unit of perceptual organization, then when the stop and fricative are in different syllables, they should be perceived as separable dimensions. Our results suggest that there is no difference at all in the perceptual processing of stops and fricatives when they are in the same syllable or different syllables. This demonstrates that the perceptual integrality we have observed holds between adjacent phonemes and does not depend on syllable structure.

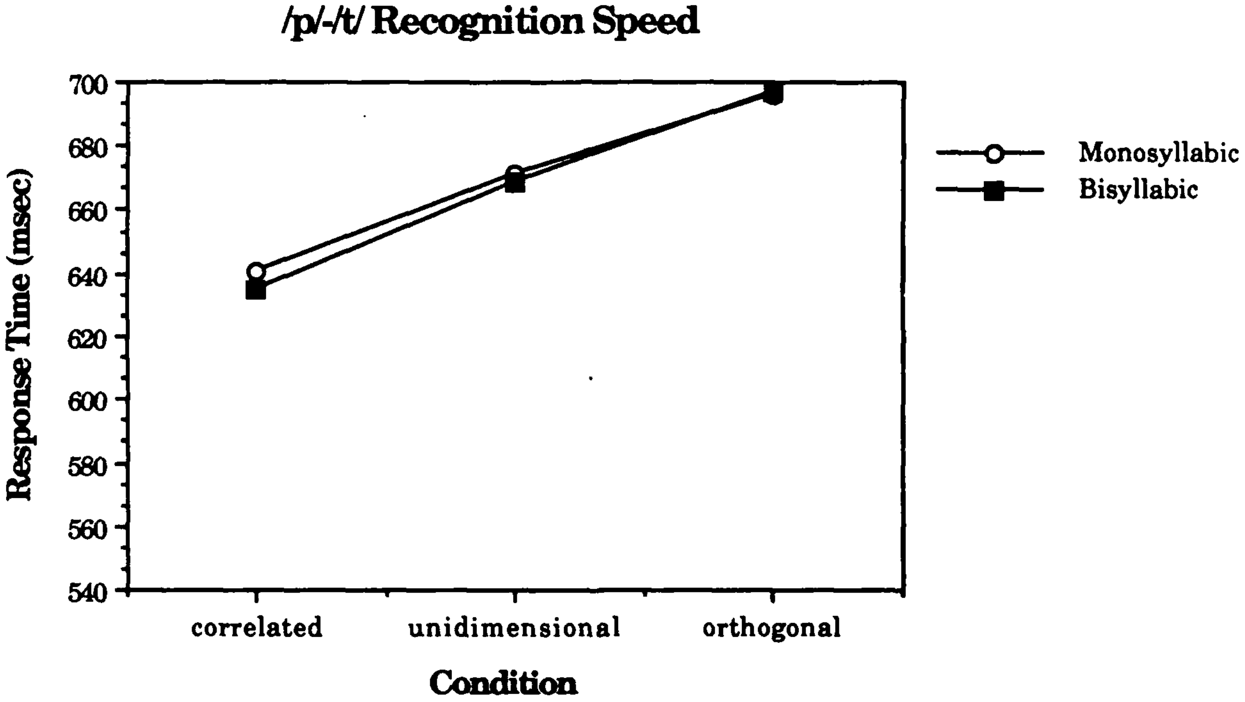

Figure 1.2. Target phoneme recognition speed in correlated, unidimensional, and orthogonal conditions when irrelevant context is varied within the same syllable as the target (circles) or in a different syllable (squares).

Ohman (1966) has demonstrated that the acoustic-phonetic effects of coarticulation span syllable boundaries in speech production so that the structure of one segment changes as the segmental context changes, even across syllables. Furthermore, coarticulation across syllable boundaries affects the listener's recognition of segmental information (e.g., Martin & Bunnell, 1982). Thus, it seems as if coarticulation in speech production is matched by a perceptual process that is informed about the distribution of acoustic information relevant to phonetic decisions across several acoustic events. In order to "decode" a phonetic segment from the speech waveform, the perceptual system must also process adjacent phonetic segments as well.

Experiment 1.2: Fricative Identification in Stop Consonant Contexts

Judgments about the identity of a stop consonant are affected by the identity of a preceding fricative, regardless of whether that fricative is in the same syllable or in a different syllable. This suggests that perceptual decoding of phonetic segments is sensitive to the coarticulatory encoding of acoustic-phonetic information into the speech waveform. However, in our first experiment, subjects always heard the fricative before the stop consonant. As a result, the subjects might find it difficult to ignore the fricative information that they had just heard when they started identifying the stop. In the present study, subjects were instructed to identify the fricative rather than the stop consonant, and we manipulated the identity of the stop consonant as the context dimension across unidimensional, correlated, and orthogonal conditions. If perceptual decoding of a target phonetic segment is dependent on using the information about adjacent, coarticulated phonemes, subjects' decisions should be affected by manipulations of the adjacent segment, even if it follows the target. The listener should wait to hear the acoustic-phonetic context before judging the target, thereby displaying the same general pattern of perceptual integrality as in the previous experiment. Of course, it is also possible that subjects may be able to judge phonemes independent of succeeding phonetic context. If this alternative account is correct, subjects' response times should be unaffected by manipulations of that context. Finally, it is possible that syllable structure could interact with the degree of foward-listening perceptual dependence, even though it did not interact with the regressive perceptual dependence in the previous experiment. As a consequence, we might find that segments in bisyllabic syllables are separable, while segments in monosyllables might be integral.

Method

Subjects. The subjects were 18 University of Chicago students and residents of Hyde Park, aged 18-31. All subjects were native speakers of English with no reported history of speech or hearing disorders. None of the subjects had participated in Experiment 1.1. The subjects were paid $4.00 an hour for their participation.

Stimuli. The stimuli consisted of the same eight monosyllabic and bisyllabic utterances from Experiment 1.1. The stimuli were presented in the same way as in the previous experiment.

Procedure. In general, the instructions, procedures, and apparatus were the same as those of Experiment 1.1, with the following exceptions. Instead of identifying the stop consonant in the stimuli, subjects were instructed to identify the fricative as /s/ or /ʃ/. For nine of the subjects, the Z key of the Macintosh Plus keyboard was labeled as the **s** response, and the / key was labeled **sh**. For the other nine subjects, the position of the **s** and **sh** labels was reversed. The choices s and **sh** were displayed on the corresponding sides of the Macintosh screen, as well.

The subjects were instructed to determine whether each stimulus contained an /s/ or an /ʃ/ sound, and to press the corresponding key as quickly as possible without sacrificing accuracy. As in Experiment 1.1, all subjects first received practice with feedback. Following the practice block of trials, all subjects received correlated, orthogonal, and unidimensional conditions which were presented without feedback. In each block, 20 repetitions of each stimulus were presented. In the correlated condition, subjects received two blocks of trials: a block of trials consisting of presentations of /spa/ and /ʃta/ was presented first, followed by a block of /sta/-/ʃpa/ trials. The unidimensional condition consisted of a /spa/-/ʃpa/ block followed by a /sta/-/ʃta/ block. The orthogonal condition was presented in a single block consisting of all four monosyllables. The order of the three conditions was counterbalanced across subjects. Each of the six possible orders was presented to three subjects.

After the monosyllables were presented, equivalent unidimensional, correlated, and orthogonal conditions were presented using bisyllabic stimuli. For example, the correlated condition consisted of an /isˈpʰa/-/iʃˈtʰa/ block and an /isˈtʰa/-/iʃˈpʰa/ block, respectively. Each subject received the bisyllabic stimulus conditions in the same order as the monosyllabic, beginning with five practice repetitions of each word, in random order. Again, each experimental block consisted of twenty repetitions of the specific stimuli for that block, in random order.

Results and Discussion

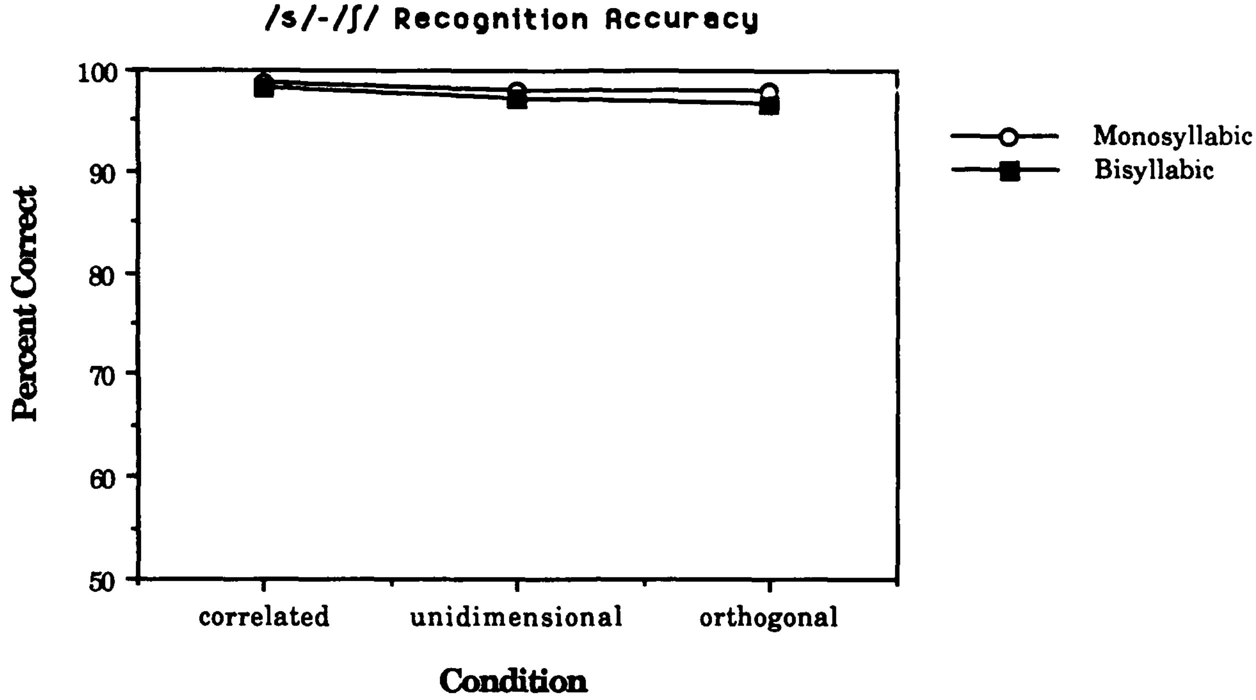

Figure 1.3. Fricative recognition accuracy when irrelevant contextual variation is within the same syllable (circles) or a different syllable (squares).

Fricative recognition accuracy was quite good, averaging over 97% correct as shown in Figure 1.3. There was no significant difference in fricative identification as a function of syllable structure (monosyllabic vs. bisyllabic), although there was a slight tendency for greater accuracy in identifying fricatives in monosyllables, $F(1, 17) = 2.51, p > .1$. There was a significant effect of condition, $F(2,34) = 4.39, p < .02$, such that accuracy was higher in the correlated condition than either the unidimensional or orthogonal conditions ($p < .05$, by post-hoc Newman-Keuls comparisons). There was no significant difference in accuracy between the unidimensional and orthogonal conditions.

Response times for fricative identification in the correlated, unidimensional, and orthogonal conditions are shown in Figure 1.4. As can be seen in this figure, there is one difference in the pattern of response times compared to the pattern observed in the previous experiment. Although subjects were faster in the unidimensional condition than in the orthogonal condition as in the first experiment, the subjects were slower in the correlated condition, which is unusual. There was no effect of syllable structure on speed of fricative identification, $F(1,17) = .564$, n.s., just as syllable structure did not affect the speed of stop classification. However, there was a significant effect of condition on fricative classification response time, $F(2,34) = 18.692, p < .001$. Response times in the unidimensional and correlated conditions were significantly faster than response times in the orthogonal condition ($\underline{p} < .05$, by Newman-Keuls comparisons). A significant interaction between syllable structure and condition, $F(2,34) = 5.092$, $p <.01$, occurred because for monosyllabic stimuli, there was no difference between response times in the unidimensional and correlated conditions, while for bisyllabic stimuli, response times in the unidimensional condition were faster than in the correlated condition ($p < .05$, by a Newman-Keuls test).

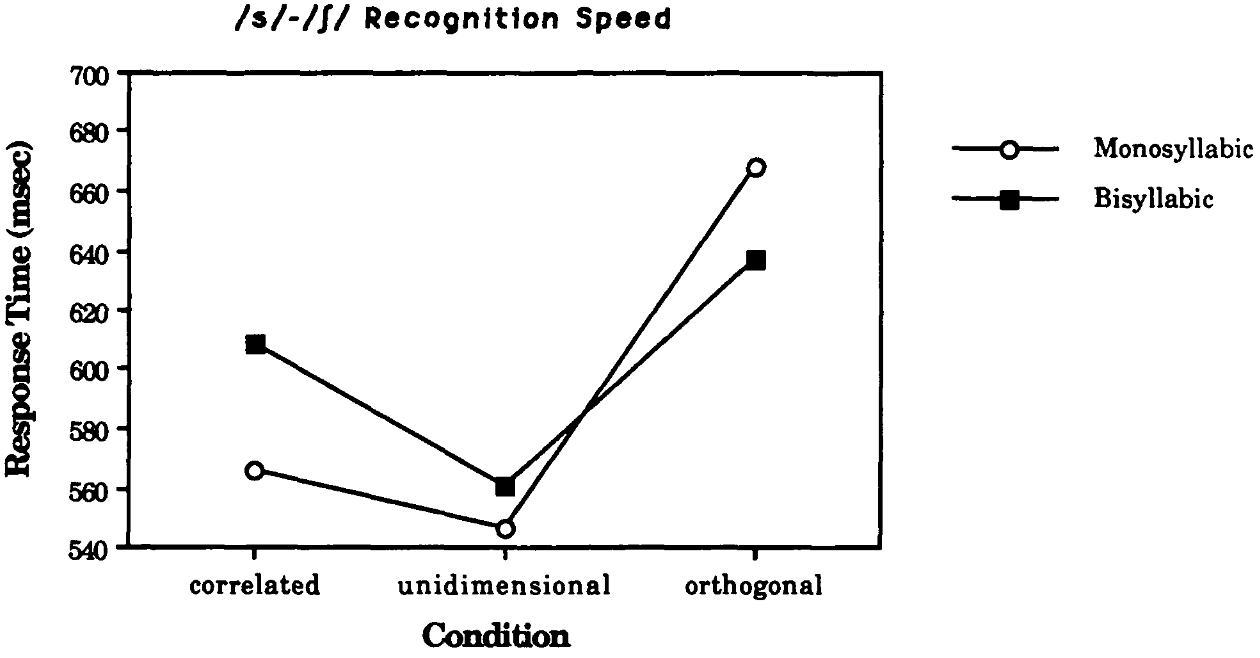

Figure 1.4. Fricative recognition times for unidimensional, correlated, and orthogonal conditions, when irrelevant contextual variation is in the same syllable (circles) and in a different syllable (squares).

The typical response time pattern revealing integrality between perceptual dimensions (Garner, 1974) is that subjects are significantly faster in the correlated condition than in the unidimensional condition. In fact, this pattern may also be observed even if the dimensions are separable, because of the ability to use the redundancy of the correlated dimension. Thus, the relative slowness of the subjects' responses in the correlated condition is somewhat surprising. However, if we consider the accuracy data together with the RTs, our results appear to be due to a speed-accuracy tradeoff between the correlated and unidimensional conditions. Subjects are significantly faster in the unidimensional condition, but they are significantly more accurate in the correlated condition.

With regard to the issue of integrality, the result of greatest importance is the finding that subjects are significantly slower to make fricative judgments in the orthogonal condition than in the unidimensional and correlated conditions. This finding parallels our results for stop consonant identification: Subjects treat adjacent phonemes as dimensions of an integral perceptual unit. The lack of any differences in integrality between monosyllabic and bisyllabic stimuli in both experiments indicates that the perceptual unit that is integral is neither the syllable nor the syllable onset. Adjacent phonemes are perceived as integral whether they are members of the same syllable or adjacent syllables.

The integrality of fricatives with adjacent stop consonants is very interesting. Remember that the fricative precedes the stop consonant in all our stimuli. When identifying the stop consonant, listeners will have already heard the most of the acoustic information corresponding to the fricative so it is not surprising that the identity of the fricative affects stop judgments. However, when the fricative is identified, subjects could potentially respond on the basis of the acoustic information preceding the stop consonant. But this doesn't seem to happen; listeners are clearly affected by the identity of the stop consonant in making their judgment of the fricative. We computed differences between response times in the orthogonal and unidimensional conditions for monosyllabic and bisyllabic stimuli for the stop consonant judgments and for the fricative judgments to determine if the increase in response times was greater for stop judgments than for fricative judgments. In other words, we examined the amount of influence of stop consonants on fricatives and fricative on stop consonants to determine whether the perceptual dependence is symmetrical or not. The difference scores were significantly greater for fricative judgments (122.9 msec for monosyllabic and 75.8 msec for bisyllabic stimuli) compared to stop judgments (38.9 msec for monosyllabic and 28.9 msec for bisyllabic stimuli, $t(34) = 2.72, p < .01$, for monosyllabic and $t(34) = 2.24, p < .05$, for bisyllabic stimuli. This indicates that fricative judgments were more dependent on stop consonants than the reverse, despite the temporal precedence of the fricatives in the utterances. Of course, we did not attempt to equate the relative discriminability of the fricatives and the stop consonants, so this asymmetry may not reflect asymmetries in integrality as much as discriminability. However, the direction of the asymmetry is interesting nonetheless.

Experiment 1.3: Consonant Identification in Vowel Contexts

In the third experiment, we investigated the integrality of consonants and vowels when the two segments occur within a single syllable and when they occur in two different syllables. We used VCV stimuli, with stress on the second vowel so that English speakers would hear the consonant as the onset of the second syllable. Subjects judged the identity of the consonant in unidimensional, correlated, and orthogonal conditions, with two sets of stimuli. In one set, we manipulated the second vowel (in the same syllable as the consonant), and in the other set of stimuli, we manipulated the first vowel.

Method

Subjects. The subjects were 24 University of Chicago students and residents of Hyde Park, aged 17 - 30. All subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hour for their participation.

Stimuli. The stimuli were 8 VCV utterances spoken by a single male talker: /o'pa/, /o'ta/, /o'pæ/, /o'tæ/, /a'po/, /a'to/, /æ'po/, and /æ'to/. In all the utterances, the second syllable was stressed, so that the second vowel would be heard as being in the same syllable as the consonant, and the syllable boundary would fall after the initial vowel. Thus, in four of the utterances the consonant was in the same syllable as the /a/ or /æ/, while in the other four the consonant and the /a-æ/ were in different syllables. These are referred to as the *within-syllable* and *between-syllable* stimuli, respectively.

Several tokens of each utterance were recorded on cassette tape in a sound-shielded booth. Digitizing, stimulus selection, and waveform editing were performed in the manner described for the first experiment. The stimuli were played to subjects over Sennheiser HD-430 headphones at approximately 79 dB SPL. Digitized stimuli were converted into speech in real time under computer control. Each waveform was played at 10 kHz through a 12-bit D/A converter and low-pass filtered at 4.6 kHz.

Procedure. The experimental procedure, apparatus, and instructions to subjects were the same as in Experiment 1.1, except as noted. Thirteen subjects had the **p** response key at their left hand and the **t** at their right; for the other eleven subjects, the position of the **p** and **t** labels was reversed. Twelve subjects heard the within-syllable stimuli first, followed by the between-syllable stimuli; twelve subjects were presented with the opposite order. Each half of the experiment began with a practice session consisting of five repetitions each of the four within-syllable stimuli or the four between-syllable stimuli. Feedback was presented as described in the first experiment.

Each block of trials consisted of 20 repetitions of each of the stimuli, presented in random order, with no feedback. In the within-syllable part of the experiment, two blocks of trials made up the correlated condition, in which variation in the stop consonant was correlated with variation in the vowel in the same syllable as the consonant. The first correlated block consisted of 20 repetitions each of /o'pa/ and /o'tæ/, and the second correlated block was composed of /o'pæ/ and /o'ta/. In the two unidimensional blocks, the stop varied while the vowel remained constant; one block consisted of 20 repetitions each of /o'pa/ and /o'ta/, and the second consisted of /o'pæ/ and /o'tæ/. The single orthogonal block consisted of 20 repetitions of each of these four stimuli.

The between-syllable portion of the experiment involved variation in a vowel that was adjacent to the stop consonant, but not in the same syllable. The correlated condition consisted of an /a'po/-/æ'to/ block and an /æ'po/-/a'to/ block. The unidimensional condition was composed of an /a'po/-/a'to/ block and an /æ'po/-/æ'to/ block. The orthogonal block included 20 repetitions of each of the four stimulus items. The sequence of unidimensional, correlated-dimension, and orthogonal-dimension blocks (within the two stimulus sets) was varied across subjects.

Results and Discussion

Figure 1.5 shows that the mean accuracy in judging the identity of the stop consonant was very high, ranging from 97% to 99% for the various conditions. There were no significant differences in accuracy among any of the conditions, or among any of the individual stimulus items.

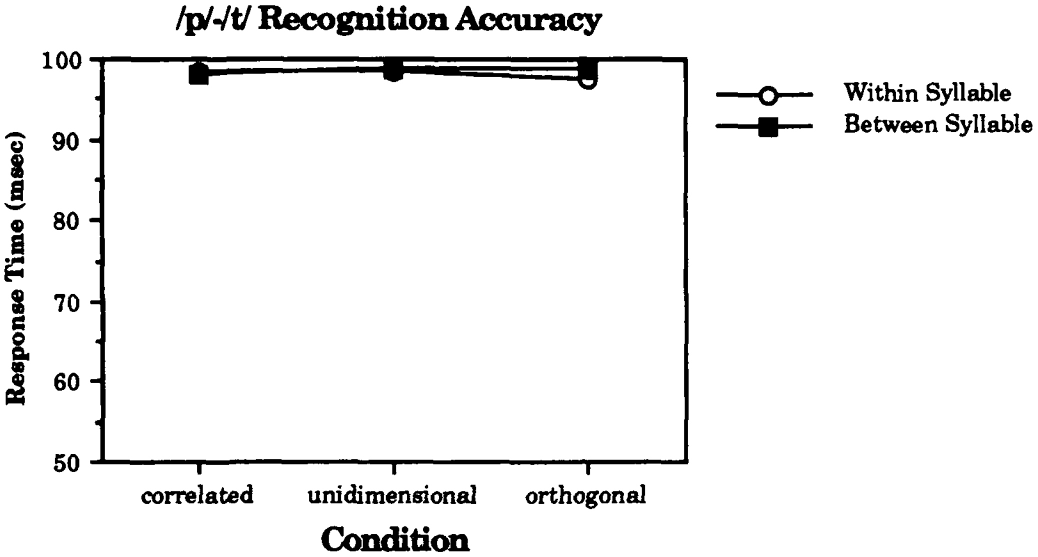

Figure 1.5. Recognition accuracy for stop consonants in unidimensional, correlated, and orthogonal conditions.

Figure 1.6 shows mean response times for stop consonant recognition for the within-syllable and between-syllable stimulus types, in the correlated, unidimensional, and orthogonal conditions. There was a significant effect of condition (correlated vs. unidimensional vs. orthogonal), $F(2,46) = 5.378$, $p < .01$. Post-hoc Newman-Keuls analyses showed that response times were significantly slower in the orthogonal condition than in the correlated condition ($p < .01$) or the unidimensional condition ($p < .05$), but that there was not a significant difference between response times in the correlated and unidimensional conditions. This follows the same overall pattern of performance as in the previous studies demonstrating that recognition of consonants depends on processing of the adjacent segments even if those segments are vowels and even if the context is in a different syllable.

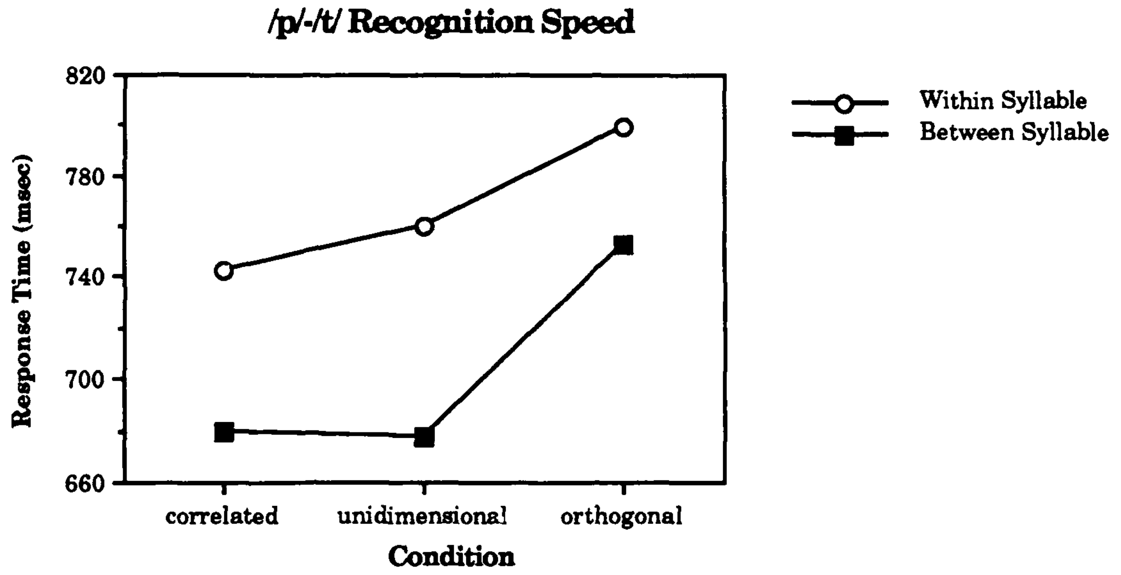

Figure 1.6. Recognition times for stop consonants in vowel context, in correlated, unidimensional, and orthogonal conditions.

Together the results of these three experiments on perceptual integrality suggest that neither the syllable nor the syllable onset are as important in the perceptual organization of speech for recognition as the segment. Furthermore, it is also clear that the phoneme is not a discrete perceptual unit. Instead, perception of a phoneme depends on recognition of adjacent phonemes as well. This perceptual effect parallels the coarticulation of segments in speech production. In speech production, the acoustic representation of a particular phoneme is affected by the production of adjacent phonemes (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). The integrality of adjacent phonemes in recognition may reflect a kind of perceptual coarticulation. Although Wood and Day (1975) were the first to demonstrate this kind of perceptual coarticulation between a consonant and vowel within a single syllable, our present findings extend this conclusion to adjacent consonants and across syllable boundaries. Just as coarticulation in speech production crosses syllable boundaries, our results suggest that perceptual coarticulation also crosses syllable boundaries and that listeners may process speech as a stream of allophonic units that are interpreted relative to the perceptual context in which they occur.

Future Studies

The results of these experiments suggest that adjacent phonemes are perceived as an integral perceptual unit, regardless of the imposed syllable structure. This suggests other experiments to explore this interpretation further. One issue that arises concerns the limits of phonetic integrality. We know that coarticulatory influences are not restricted to immediately adjacent phonemes. For example, the /u/ in /stru/ affects the /s/ differently from the /i/ in /stri/. Given that adjacent phonetic segments are perceptually integral, how far along a phonetic sequence does this integrality extend? Do phonemes that are separated by another segment show this same degree of integrality or does integrality between segments drop off with ordinal separation? The perceptual representation of speech may be allophonic incorporating aspects of immediately preceding and succeeding segments or this representation may extend over a much broader span of context.

We have investigated the integrality of syllables and found no special perceptual status conferred by syllable membership. However, it seems reasonable to ask whether other, higher-level linguistic units are perceived as integral. For example, spoken words might be processed as integral perceptual units. Thus, the goal of a second study will be to determine whether a decision about a target phoneme in one word is affected less by changes in a context phoneme in a second, adjacent word, compared to changes in the same context phoneme when it occurs in the same word as the target. For example, subjects could judge whether the following sequences contain /r/ or /l/ for unidimensional, orthogonal, and correlated conditions for within and between word stimulus sets. Within word a unidimensional condition might be *row broom* vs. *row bloom* and a correlated condition might be *row broom* vs. *row gloom* and the orthogonal condition would consist of all /b/ and /g/ combinations with /l/ and /r/. Between words, a unidimensional condition would place the stop consonant in the previous word such as *robe room* vs. *robe loom* and the correlated condition would consist of *robe room* vs. *rogue loom* with the orthogonal condition including all four stimuli. A set of nonword control conditions will also be constructed to match these word conditions.

IV. Project 2: Capacity Demands of Talker Normalization

Talkers differ in the size and length of their vocal tracts. As a result, the acoustic structure of vowels produced by different talkers may be extremely different. Two talkers may produce the same intended vowel such as /a/ (as in hot) with very different pattern structures and they may produce different vowels such as /a/ and /^/ (as in hut) with the same pattern structure (Petersen & Barney, 1952). In order to recognize any vowel produced by a talker, the listener must know something about the structure of the set of vowels produced by that talker in order to correctly interpret the acoustic cues.

When all the vowels produced by a single talker are plotted in a space defined by the frequencies of the first and second formants (F1 and F2), these vowels are arrayed in a roughly triangular region with /i/, /a/, and /u/ (also called the point vowels) as the vertices of the space. The vowel spaces for different talkers are typically nonlinear transforms of each other, so that normalization of talker differences is not a simple scaling operation (Fant, 1973; Morin & Nusbaum, 1988).

Two different mechanisms have been described for carrying out the process of normalizing talker differences. Contextual tuning uses samples of vowels produced by a talker to map out a representation of the talker's vowel space (cf. Gerstman, 1968; Sawusch, Nusbaum, & Schwab, 1980). Once a representation of the vowel space is constructed, any acoustic vowel token can be mapped directly to the correct region of phonetic space.

Structural estimation uses information contained within a *single* vowel token to normalize talker differences. Syrdal and Gopal (1986) have shown that pitch information and formants above F2 provide a sort of relative framework within which F1 and F2 can be recognized, although not perfectly. Thus, structural estimation does not need to sample any more speech than the token that must be recognized.

Verbrugge and Rakerd (1986) have suggested that the dynamic specification of vowels by the consonant transitions in CVC syllables may provide another source of information for resolving talker differences within a single token. Thus, there have been proposed two different forms of structural estimation. One is based on static properties of the vowel spectrum, while the other is based on the dynamic properties of coarticulatory information.

Experiment 2.1: Normalization of Isolated Vowels and CVCs

To investigate the operation of contextual tuning and structural estimation, we carried out a vowel monitoring experiment. The task was quite simple. Subjects were told to listen for a target vowel such as "EE as in BEAT" in a sequence of utterances and they are told to press a button quickly and accurately for every recognized occurrence of the target. In one condition (the blocked-by-talker condition), in each trial, all the utterances were produced by a single talker. Across different blocks of trials, subjects monitored for vowels produced by four different talkers. In a second condition (the mixed-talker condition), within each trial, the utterances were produced by a mix of the four different talkers. Thus, in the blocked condition, contextual tuning could operate to resolve talker differences since listeners only heard vowels from one talker at a time, whereas in the mixed condition, only structural estimation could operate.

If recognition performance is the same in the blocked and mixed conditions, this would provide evidence for the operation of structural estimation. If recognition performance is significantly worse in the mixed condition, this would provide evidence for contextual tuning. Moreover, one group of subjects monitored for isolated vowels, while the remainder monitored for vowels in CVCs. If dynamic specification of vowel identity is necessary for structural estimation, there should be no difference in performance between blocked and mixed conditions for CVCs, but a large difference for isolated vowels.

Method

Subjects. The subjects were 22 University of Chicago students and Hyde Park residents. Each subject participated in a single hour-long session. All subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hour for their participation.

Stimuli. Two sets of stimuli were used in this experiment. The first set consisted of the eight isolated vowels /i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /u/, /ʊ/, and /ʌ/. The second set consisted of the same eight vowels produced as CVC syllables with the consonant frame /rVk/. All stimuli were spoken by two male and two female talkers. The stimuli were recorded on cassette audiotape. The recorded utterances were then digitized at 10 kHz using a 12-bit A/D converter after low-pass filtering at 4.6 kHz. The waveforms were edited into separate stimulus files using a digital waveform editor with 100 microsec accuracy. The stimuli were edited so that each waveform began with the first glottal pulse of the utterance.

The stimuli were converted to analog form by an IBM-PC/AT at 10 kHz using a 12-bit D/A converter and were low-pass filtered at 4.6 kHz. The stimuli were played binaurally to listeners over Sennheiser HD-430 headphones at approximately 76 dB SPL.

Procedure. Experimental sessions were carried out with one to three subjects per session. The subjects were randomly assigned to two groups of 11 subjects each. One group was presented with the CVC stimuli, while the other group heard only the isolated vowels. The task was to monitor a sequence of 16 vowels or syllables for the occurrence of a designated target vowel.

All subjects participated in two conditions. In one condition, trials were blocked by voice so that all the stimuli for each trial were produced by a single talker. In this condition, the subjects received eight trials for each of the four talkers, one talker after another. The order of the talkers was randomly determined for each experimental session. In the second condition, each trial consisted of stimuli produced by all four talkers, so that the stimuli were mixed across talkers within every trial.

Each trial consisted of a sequence of 16 stimuli, each stimulus separated by a 250 msec interstimulus interval. Subjects were seated in front of a Macintosh computer and their task was to press a button on the keyboard as quickly and as accurately as possible every time a designated stimulus target was heard. Four occurrences of a single target were presented at random positions on every trial, with no target presented as the first or last stimulus in a trial, or immediately following a previous occurrence of a target. Each trial began with a short beep sound produced as a warning signal by the computer with the word READY appearing on the computer screen for three sec. Following the ready signal, the target vowel for that trial was displayed on the screen in the form "OO as in BOOK." After another three sec interval, a sequence of stimuli was presented over headphones and the subjects' responses were collected and stored by the computer. After all 16 stimuli for the trial were presented, the beep and READY signal were presented again signalling the beginning of the next trial.

The subjects were given three practice trials in the blocked condition to familiarize them with the trial structure and task. Following practice, subjects received four blocks of eight trials each, one block for each of the four talkers. Each block consisted of two trials with each of the target vowels /i/, /I/, /u/, and /U/ (isolated vowel group) or target CVCs /rik/, /rIk/, /ruk/, and /rUk/ (CVC group). The sequence of eight trials in each block was randomly determined for each session.

The mixed condition was very similar to the blocked condition with the following exceptions. Each trial included distractors and targets from each of the four talkers, with one target occurrence from each talker making up the four target occurrences for a trial. Subjects were instructed to respond to the indicated target if it was spoken by any of the talkers. Following three practice trials, the subjects received four blocks of eight trials each, with each block again consisting of two trials with each of the four targets. The order of trials was randomly determined for each session and the order of conditions (blocked and mixed) was counterbalanced across subjects.

Results and Discussion

There are two basic issues regarding vowel normalization that this experiment addresses. First, two mechanisms have been proposed to mediate normalization of talker differences: contextual tuning and structural estimation. In the blocked-talker condition, listeners can use the contextual tuning mechanism since they are only listening to one talker at a time. In the mixed talker condition, the talker may change from stimulus to stimulus within a trial, so contextual tuning will not work. If performance is better in the blocked-talker condition than the mixed-talker condition, this would provide support for the operation of a normalization mechanism that uses several tokens of a talker's vowel space (contextual tuning). If listeners are completely unable to recognize vowels in the mixed-talker condition, this would suggest that listeners can only rely on contextual tuning for normalization. On the other hand, if performance is equally good in the blocked and mixed conditions, this would suggest that listeners need only use the information contained within a single vowel token for normalization of talker differences. Second, if listeners use the dynamic specification of a vowel by consonant transitions to normalize talker differences, then any differences between blocked and mixed conditions should be reduced for CVC syllables compared to isolated vowels.

Three measures of vowel recognition performance were analyzed for our monitoring task: percentage of correct detections (hits), response times (RT) for hits, and percentage of false alarms. Response times were measured from the onset of each stimulus presentation within a trial. Response times less than 150 msec were attributed to the immediately preceding stimulus. Thus, the response time for the previous stimulus was computed as the duration of the preceding stimulus plus interstimulus interval plus the recorded response time.

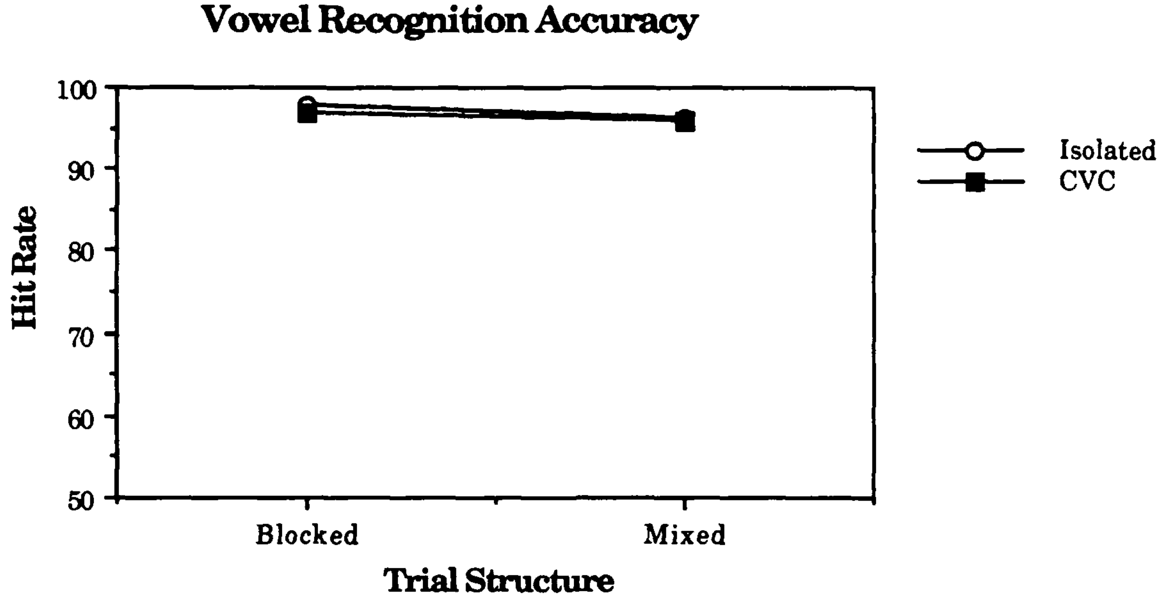

Figure 2.1. Mean correct vowel target recognition in trials with only one talker (blocked) or a mix of four talkers (mixed).

Figure 2.1 shows the mean hit rate for the isolated vowel and CVC groups for the blocked-talker and mixed-talker conditions. Performance is generally quite good across conditions, typically exceeding 95% correct responses. Although the difference in hit rate between the blocked (97% correct) and mixed (96% correct) conditions is quite small, subjects were significantly more accurate in the blocked-talker condition, $F(1, 20) = 7.56, p < .02$. This suggests that listeners may indeed use contextual tuning for talker normalization. However, the high level of performance for the mixed condition indicates that listeners can also use structural estimation for normalization. The lack of a significant difference between performance on isolated vowels and CVCs, F(1,20) = .216, n.s., and the lack of an interaction between stimulus type (isolated vowels vs. CVCs) and condition (blocked vs. mixed), $F(1,20) = .140$, n.s., suggests that the consonant transitions may provide little, if any advantage in vowel recognition. Of course, the high recognition rates may obscure any differences between isolated vowels and CVCs.

Figure 2.2 displays the mean false-alarm (FA) rate for the isolated vowel and CVC groups in the blocked and mixed conditions. Although the CVC group showed significantly higher FA rates, $F(1, 20) = 6.30, p < .03$, than the isolated vowel group, both group's FA rates were below 3% and there was no significant interaction between stimulus type (isolated vowels vs. CVCs) and condition (blocked vs. mixed). Although there was no significant difference in FA rates in the blocked and mixed conditions, the results argue against any facilitation of vowel recognition by the consonant frame in the CVCs. Furthermore, considering the hit and FA data together suggests that changes in vowel perception due to differences in the blocked and mixed conditions are due to greater perceptual sensitivity in the blocked-talker condition.

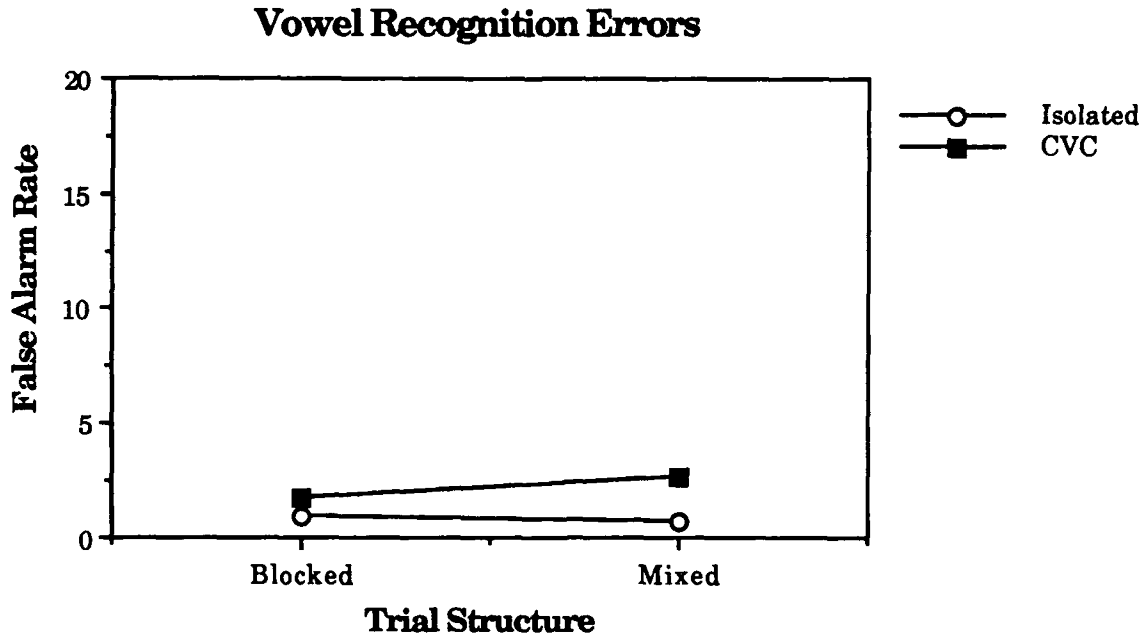

Figure 2.2. False alarms in vowel monitoring when subjects listened to one talker at a time (blocked) or a mix of four different talkers (mixed).

Figure 2.3 shows the mean response times for the isolated vowel and CVC groups in the blocked-talker and mixed-talker conditions. Response times in the mixed condition were about 28 msec longer than response times in the blocked condition, $F(1, 20) = 14.80$, $p < .001$. This provides further evidence that the process of recognizing vowels is impaired by the absence of contextual tuning. In addition, response times were about 70 msec longer for subjects monitoring for CVCs than the response times for subjects monitoring for isolated vowels, $F(1, 20) = 11.73$, $p < .003$. This difference may simply reflect the duration of the transitions for the /r/ at the beginning of the CVCs. More important is the lack of a significant interaction between stimulus type and condition, $F(1,20) = .005$, n.s., indicating that the increases in response times for the mixed condition relative to the blocked condition were almost identical for the CVC and isolated vowels groups. The CVCs do not appear to provide any special normalization advantage over isolated vowels in the mixed condition.

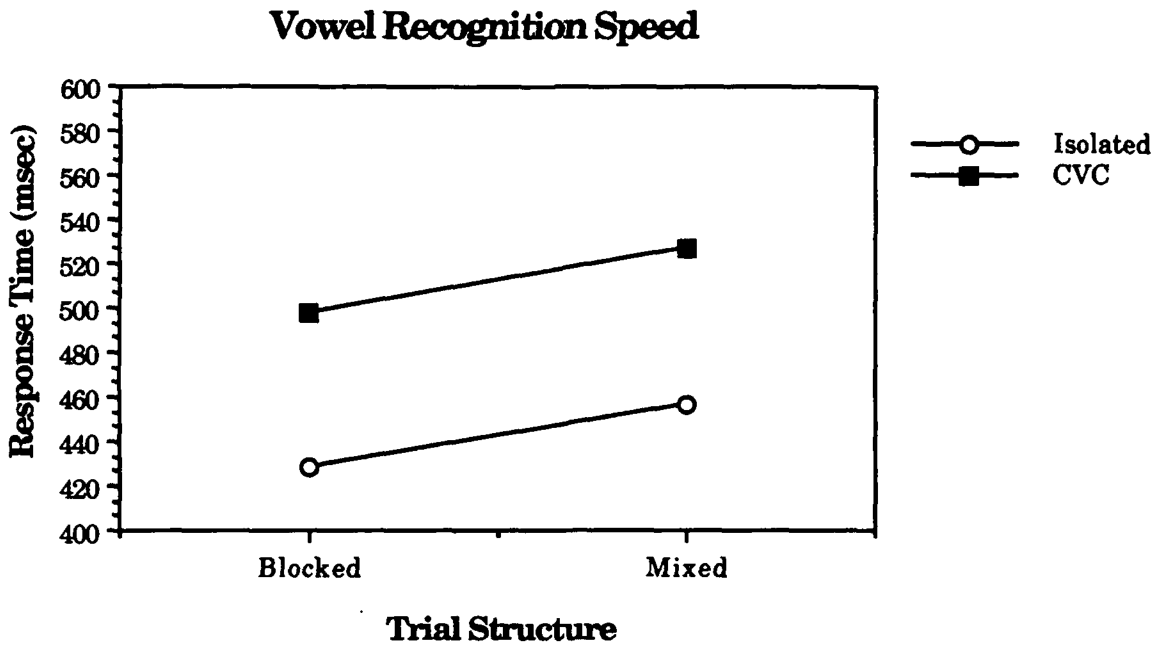

Figure 2.3. Vowel recognition time for hits when each trial consists of speech from one talker at a time (blocked) or a mix of four talkers (mixed).

If listeners normalize talker differences in vowel perception using only the information contained within a single vowel token (e.g., Syrdal & Gopal, 1986), there should be no difference in performance between the blocked-talker and mixed-talker conditions. However, we found significantly better accuracy and faster response times for vowel recognition in the blocked-talker condition compared to the mixed-talker condition. Listeners are using the information about a talker that is gathered from a collection of speech tokens in the blocked condition to recognize vowels faster and more accurately. This suggests that listeners are recognizing vowels using a mechanism like contextual tuning by which some representation of a talker's vowel space is constructed as a reference for recognition. This finding argues against the prior claims of Verbrugge and Rakerd (1986). At the same time, it is important to note that the performance differences between the two conditions are small, albeit reliable. Therefore, it is clear that listeners do not just use contextual tuning, but are also able to use the information within a single vowel token to normalize talker differences as well. It appears as though this structural estimation mechanism may be less accurate and may either be slower, or require more effort. Thus, our results provide the first evidence that listeners may use both mechanisms to normalize talker differences in vowel perception. Finally, we found no evidence to support the claims that dynamic specification of vowels confers special advantage in vowel perception or for talker normalization, contrary to several recent claims (e.g. Verbrugge & Rakerd, 1986). Consonant transitions may provide information about vowel identity under some conditions, but they did not reduce the effort required by listeners to normalize talker differences.

Experiment 2.2: Normalization of Consonants

The results of our first experiment on normalization of talker differences indicate that listeners use both structural estimation and contextual tuning mechanisms in vowel recognition. However, Rand (1971) demonstrated that the placement of category boundaries between consonants differing in place of articulation is dependent on the vocal tract characteristics of the talker. His results do not, however, address the issue of what the mechanisms underlying this consonant normalization effect might be. In an effort to address this question, the present study investigates the normalization of consonants using the same target monitoring paradigm used in Experiment 2.1.

Method

Subjects. The subjects were 12 University of Chicago students and Hyde Park residents. Each subject participated in a single hour-long session. All subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hour for their participation.

Stimuli. The stimuli consisted of a set of eight consonant-vowel syllables: /da/, /ta/, /ga/, /ka/, /ba/, /pa/, /ma/, and /na/. All stimuli were spoken by two male and two female talkers. The stimuli were presented to listeners in real-time under control of an IBM-PC/AT computer as described in the previous study.

Procedure. Experimental sessions were carried out with one to three subjects per session. All subjects participated in two conditions. In one condition, trials were blocked by voice so that all the stimuli for each trial were produced by a single talker. In this condition, the subjects received eight trials for each of the four talkers, one talker after another. The order of the talkers was randomly determined for each experimental session. In the second condition, each trial consisted of stimuli produced by all four talkers, so that the stimuli were mixed across talkers within every trial. The order of conditions was counterbalanced across subjects.

The subjects were given three practice trials in each condition to familiarize them with the trial structure and task. Following practice, subjects received four blocks of eight trials each, with each block consisting of two trials with each of the target consonants /da/, /ta/, /ba/, /pa/. The sequence of eight trials in each block was randomly determined for each session. For the blocked-by-talker condition, subjects received one block for each of the four talkers; for the mixed-talker condition, subjects received the same number of blocks and trials, but the stimuli for each trial were drawn from the set of all four talkers. Thus, the only difference between the blocked and mixed talker conditions was the arrangement of stimuli during trials.

Each trial consisted of a sequence of 16 stimuli, each stimulus separated by a 250 msec interstimulus interval. Subjects were seated in front of a Macintosh computer and their task was to press a button on the keyboard as quickly and as accurately as possible every time a designated stimulus target was heard. Four occurrences of a single target were presented at random positions on every trial, with no target presented as the first or last stimulus in a trial, or immediately following a previous occurrence of a target. Each trial began with a short beep sound produced as a warning signal by the computer with the word READY appearing on the computer screen for three sec. Following the ready signal, the target consonant for that trial was displayed on the screen in the form "b as in bee." After another three sec interval, a sequence of stimuli was presented over headphones and the subjects' responses were collected and stored by the computer. After all 16 stimuli for the trial were presented, the beep and READY signal were presented again signalling the beginning of the next trial.

Results and Discussion

This experiment addresses the basic issue of what mechanisms underlie the perceptual normalization of talker differences. As in the first experiment, listeners can use the contextual tuning mechanism in the blocked-talker condition since they are only listening to one talker at a time. In the mixed talker condition, since the talker may change from stimulus to stimulus within a trial, contextual tuning will not work. Thus, if subjects perform better in the blocked-talker condition than the mixed-talker condition, this would provide support for the operation of a contextual tuning normalization mechanism. On the other hand, if performance is equally good in the blocked and mixed conditions, this would suggest that listeners need only the information contained within a single CV token to normalize talker differences.

Three measures of consonant recognition performance were computed for the monitoring task in this experiment: percentage of correct detections (hits), response times (RT) for hits, and percentage of false alarms. Response times were measured from the onset of each stimulus presentation within a trial. Response times less than 150 msec were attributed to the immediately preceding stimulus; the response time for the previous stimulus was computed as the duration of the preceding stimulus plus interstimulus interval plus the recorded response time.

Figure 2.4 shows that the mean hit rate for the CV syllables in both the blocked (98.8%) and the mixed (99.1%) groups was quite high. Taken alone, the lack of a difference between the groups, F(1,11) = .449, might seem evidence for the operation of only structural estimation. There appears to be no improvement in performance even when consistent information about a talkers vocal characteristics is present in the blocked-by-talker condition. It is perhaps more likely, however, that the high recognition rates obscure any differences between the blocked-by-talker and mixed-talker condition.

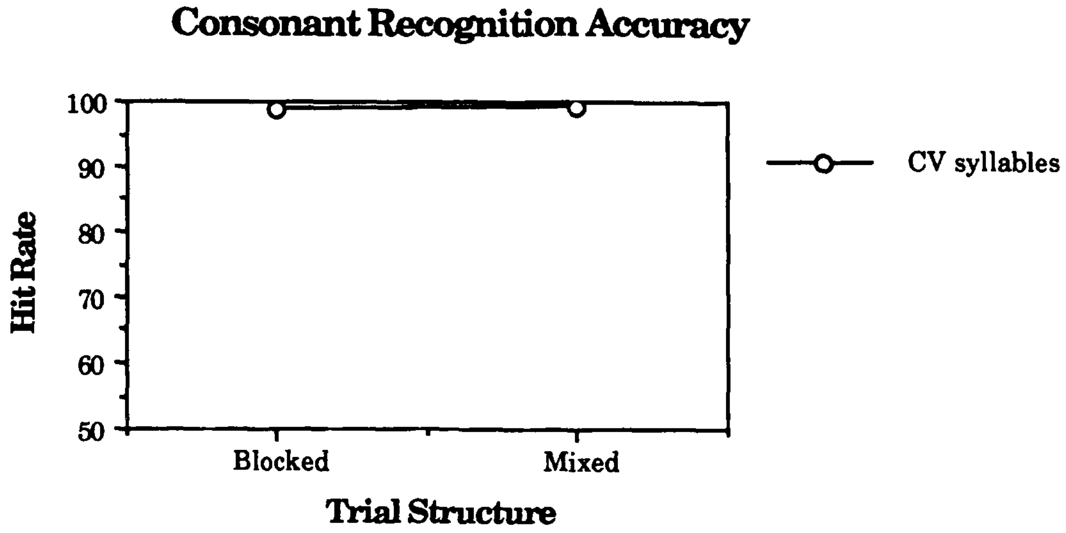

Figure 2.4. Mean correct consonant target recognition in trials with only one talker (blocked) or a mix of four talkers (mixed).

Similarly, the false alarm rates for the blocked-by-talker (.56%) and mixed-talker (.67%) conditions plotted in Figure 2.5 demonstrate no significant difference; F(1,11) = .376. Again, however, the high accuracy of performance may obscure any differences between the two conditions.

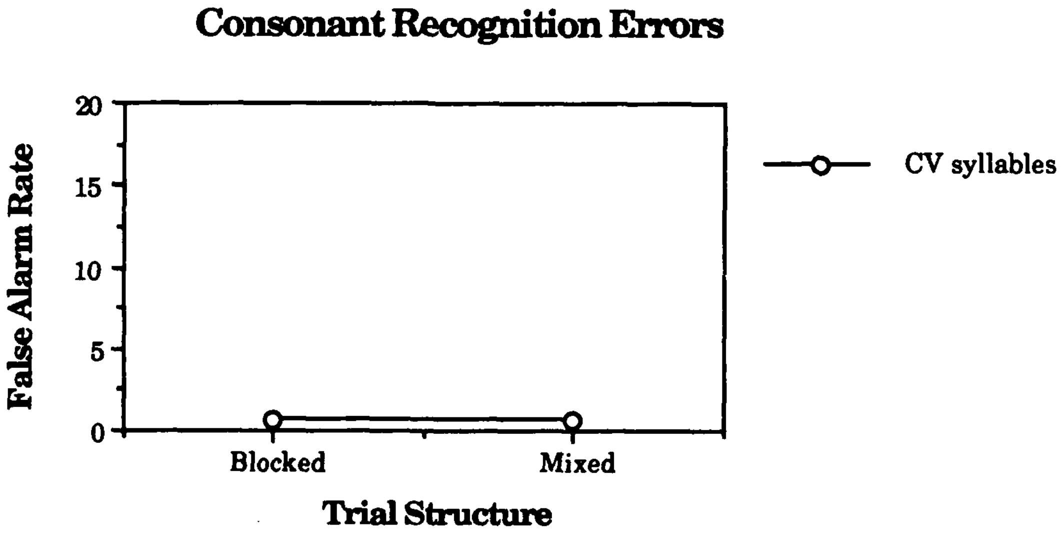

Figure 2.5. False alarms in consonant monitoring when subjects listened to one talker at a time (blocked) or a mix of four different talkers (mixed).

Figure 2.6, on the other hand, shows that the mean response time for the the mixed-talker condition is about 13 msec slower than in the blocked-by-talker condition, F(1,11) = 5.1, p < .05. This provides evidence that the process of recognizing consonants produced by different talkers may indeed involve the use of contextual tuning mechanisms. The slower response times suggest that recognition in the mixed-talker condition may require more attention and effort than in the blocked-talker condition.

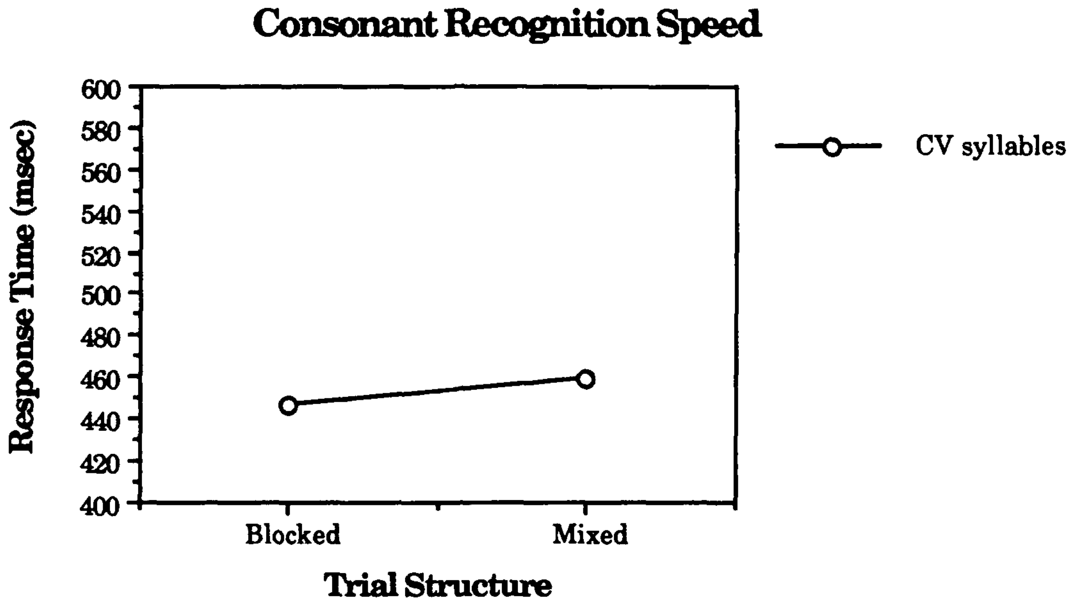

Figure 2.6. Consonant recognition time for hits when each trial consists of speech from one talker at a time (blocked) or a mix of four talkers (mixed).

The high hit rate and low false alarm rate in both the mixed-talker and blocked-talker conditions provides clear evidence that listeners do not just use contextual tuning to recognize consonants spoken by different talkers, but are also able to use the information within a single CV token to normalize these differences as well. However, significantly faster response times for consonant recognition in the blocked-talker condition compared to the mixed-talker condition indicates that listeners are using information gathered about a specific talker to aid in their recognition of consonants. This suggests that listeners are recognizing consonants using a mechanism like contextual tuning by which some representation of a talker's vocal characteristics are used as a reference for recognition. Although the exact nature of the information that is used by the listener to track or map a particular talker remains to be specified, it appears that its operation is similar to that demonstrated by vowel tokens. Although Syrdal and Gopal's (1986) model sets forth what this information might be for vowels, their treatment cannot be directly applied to the quickly changing frequency characteristics of stop consonants. Clearly, there is a need for a more general model of talker normalization.

Experiment 2.3: Normalization of Words

Our results from the previous two experiments suggest the operation of two different normalization mechanisms in recognition of vowel information in isolation and in CVC contexts, and in recognition of consonant information in CV context. The contextual tuning mechanism normalizes talker differences based on processing several vowel or consonant tokens from the same talker. The structural estimation mechanism normalizes talker differences based on the information contained within a single token, although this requires more effort and attention. It can be argued, however, that in understanding spoken language, word recognition is much more critical than consonant or vowel recognition in the context of nonsense syllables. Perhaps in the recognition of spoken words, these normalization effects are greatly overshadowed by the linguistic redundancy inherent in spoken language, which may reduce the capacity demands imposed by talker normalization. On the other hand, if the same type of normalization effect is found for recognition of spoken words as found for phonemes, this would suggest that low-level acoustic-phonetic recognition processes may provide a fundamental limit on speech comprehension. Although a recent study by Mullennix, Pisoni, and Martin (1989) suggests that normalization may be required for spoken words, it does not suggest mechanisms by which this may occur. The present study extends the target monitoring task used in the previous two experiments to investigate the roles of structural estimation and contextual tuning in the normalization of spoken words.

Method

Subjects. The subjects were 8 University of Chicago students and Hyde Park residents. Each subject participated in a single hour-long session. All subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hour for their participation.

Stimuli. The stimuli consisted of a set of nineteen phonetically balanced words: *ball*, *tile*, *cave*, *done*, *dime*, *cling*, *priest*, *lash*, *romp*, *knife*, *reek*, *depth*, *park*, *gnash*, *greet*, *jaw*, *jolt*, *bluff*, and *cad*. All stimuli were spoken by two male and two female talkers, and were digitized, filtered, and editing as described in Experiment 1. The stimuli were presented to listeners in real-time under control of an IBM-PC/AT computer as described in the previous studies.

Procedure. Experimental sessions were carried out with one to three subjects per session. All subjects participated in two conditions. In one condition, subjects listened for target words in spoken sequences of phonetically-balanced words produced by a single talker. Following the set of trials for one talker, the subjects then heard another series of trials with all of the PB words produced by a different talker. In this manner, subjects listened to words produced by each of the four talkers. The order of the talkers was randomly determined for different experimental sessions under computer control. In the other condition, subjects listened for target words in spoken sequences produced by a mix of four different talkers. In both conditions, the task was to monitor a sequence of 16 words for the occurrence of a designated target word. The order of conditions was counterbalanced across subjects.

The subjects were given three practice trials in each condition to familiarize them with the trial structure and task. Following practice, subjects received four blocks of eight trials each, with each block consisting of two trials with each of the target words *ball*, *tile*, *cave*, *done*. These word targets differ from the distractors in several phonemes so that no minimal pairs are formed. The sequence of eight trials in each block was randomly determined for each session. In the blocked-by-talker condition, subjects received one block for each of the four talkers; in the mixed-talker condition, subjects received the same number of blocks and trials, but the stimuli for each trial were drawn from the set of all four talkers. Thus, the same word targets and distractors and talkers were used in each condition. The only difference was the arrangement of stimuli during trials.

Each trial consisted of a sequence of 16 stimuli, each stimulus separated by a 250 msec interstimulus interval. Subjects were seated in front of a Macintosh computer and their task was to press a button on the keyboard as quickly and as accurately as possible every time a designated stimulus target was heard. Four occurrences of a single target were presented at random positions on every trial, with no target presented as the first or last stimulus in a trial, or immediately following a previous occurrence of a target. Each trial began with a short beep sound produced as a warning signal by the computer with the word READY appearing on the computer screen for three seconds. Following the ready signal, the target word for that trial was displayed on the screen in the form "ball." After another three second interval, a sequence of stimuli was presented over headphones and the subjects' responses were collected and stored by the computer. After all 16 stimuli for the trial were presented, the beep and READY signal were presented again signalling the beginning of the next trial.

Results and Discussion

This experiment addresses the question of whether high level lexical knowledge that is brought to bear on a word recognition task can override the perceptual normalization process. If this were the case, we would expect no difference between the blocked-talker and mixed-talker conditions. If differences do exist, however, this would suggest that the same mechanisms that underlie the perception of vowels and consonants also apply to words, despite the activation of lexical information. If subjects perform better in the blocked-talker condition than the mixed-talker condition, this would provide support for the operation of a contextual tuning normalization mechanism. On the other hand, if performance is equally good in the blocked and mixed conditions, this would suggest that listeners need only the information contained within a single word token to normalize talker differences.

Three measures of word recognition performance were computed for the monitoring task in this experiment: percentage of correct detections (hits), response times (RT) for hits, and percentage of false alarms. Response times were measured from the onset of each stimulus presentation within a trial. Response times less than 150 msec were attributed to the immediately preceding stimulus; the response time for the previous stimulus was computed as the duration of the preceding stimulus plus interstimulus interval plus the recorded response time.

For spoken words, the pattern of hits and false alarms was quite similar to the results observed in vowel and consonant recognition. The high hit rates and low false alarm rates for both the blocked (hits: 98.0%; false alarms: 1.06%) and mixed (hits: 96.6%; false alarms: 1.03%) groups, and the lack of a difference between the two, F(1,7) = .761 for hits, F(1,7) = .003 for false alarms, suggests the operation of a structural estimation mechanism. There appears to be no improvement in performance even when consistent information about a talker's vocal characteristics is present in the blocked-by-talker condition.

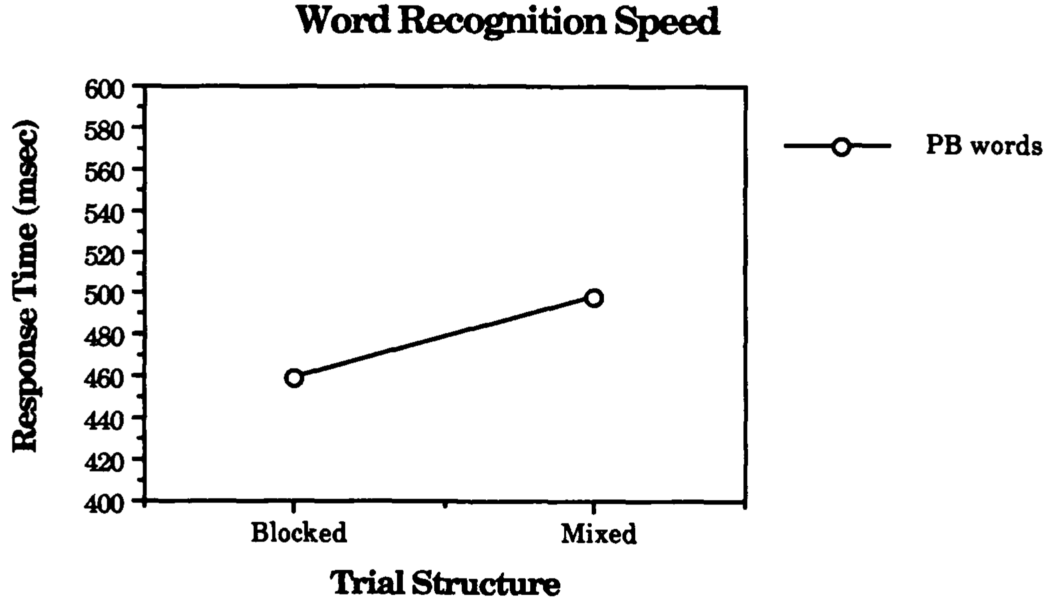

Figure 2.7. Word recognition time for hits when each trial consists of speech from one talker at a time (blocked) or a mix of four different talkers (mixed).

The response times, however, again provide the most interesting information about perceptual processing of talker vocal characteristics. Figure 2.7 shows that the mean response time for the the mixed-talker condition is about 39 msec slower than in the blocked-by-talker condition, $F(1,7) = 8.9, p < .03$. This provides evidence that the process of recognizing words produced by different talkers may indeed involve the use of contextual tuning mechanisms.

The high accuracy rate and low error rate in both the mixed-talker and blocked-talker conditions provides clear evidence that listeners do not just use contextual tuning to recognize words spoken by different talkers, but are able to use the information in a single word token to normalize these differences as well. However, significantly faster response times for word recognition in the blocked-talker condition compared to the mixed-talker condition indicates that listeners are using information gathered about a specific talker to aid in their recognition of words. As with consonants, present normalization models (e.g., Syrdal &

Gopal, 1986) were proposed to account only for the normalization of vowels and are therefore insufficient to account for the present findings. In any event, the evidence provided by the present study, that some form of contextual tuning must take place for word recognition, suggests that low-level acoustic-phonetic recognition processes provide a fundamental limit on speech comprehension.

Experiment 2.4: Effects of Memory Preload on Vowel and Consonant Normalization

The results of our first three experiments on normalization of talker differences indicate that listeners use both structural estimation and contextual tuning mechanisms in vowel recognition. Taken alone, however, the results of these studies do not indicate the reason that subjects are slower in recognizing target tokens based on structural estimation alone in the mixed condition. One possibility is that listeners may use structural estimation to learn about characteristics of a talker and if those characteristics are consistent with the next token of speech, they can be used to facilitate recognition. Otherwise, if the next token of speech is produced by a different talker, the listener will have to compute the vocal characteristics of the talker again. If the listener is slower to recognize speech in the mixed condition because of such a relearning process, then this suggests that structural estimation may require more effort and attention than contextual tuning. Once contextual tuning establishes a representation of a talker's vowel space, recognition of any particular vowel token may be a relatively simple mapping operation from the acoustic pattern of the stimulus to a region of the vowel space. However, for structural estimation, the normalization process and recognition process must proceed in real-time and may impose capacity demands.

To test this hypothesis, we gave subjects the same task that subjects performed in the first three studies. In addition to this task, though, subjects were given a secondary task to perform that would demand attention, thereby reducing the cognitive capacity available for speech perception. For the secondary task, subjects were given a list of numbers to remember throughout the course of the target monitoring task. By varying the length of the list, we can vary the demands placed on working memory. If increased demands on working memory (Baddeley, 1986) have an impact on target recognition time, this would suggest that target recognition is limited by central capacity. One group of subjects received trials that were blocked by talker and a second group received trials in which the vowels were produced by a mix of four different talkers. Differences in the effects of number preload (i.e. availability of working memory) on monitoring times in the blocked and mixed conditions should indicate the demands of contextual tuning and structural estimation on central capacity.

Method

Subjects. The subjects were 50 University of Chicago students and Hyde Park residents. Each subject participated in a single hour-long session. All subjects were native speakers of English with no reported history of speech or hearing disorders. The subjects were paid $4.00 an hour for their participation.

Stimuli. The stimuli consisted of the set of eight isolated vowels used in Experiment 1 (/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /u/, /ʊ/, and /ʌ/) and the set of eight consonant-vowel syllables used in Experiment 2 (/da/, /ta/, /ga/, /ka/, /ba/, /pa/, /ma/, and /na/). All stimuli were produced by two male and two female talkers. The stimuli were presented to listeners in real-time under control of an IBM-PC/AT computer as described in the previous studies.

Procedure. The subjects were seated in front of Macintosh computers which collected responses and response times. They were tested in small groups of one to three subjects each. Eighteen of the subjects were assigned to listen to isolated vowels and monitor for specific vowel targets, and the remainder of the subjects were assigned to listen to CV syllables and monitor for consonant targets. About half of the subjects from each of these two groups were then assigned to either of two other groups, the blocked-by-talker group and the mixed-talker group. In all groups, on each trial, subjects were given a visually presented list of numbers to remember. Following presentation of the number preload list, subjects monitored a spoken sequence of tokens for a designated target token. After the monitoring task, subjects were prompted to recall the visually presented numbers in their original order. In one condition, the preload list consisted of a single two-digit number, whereas in a second condition, the preload consisted of three two-digit numbers. The order of the preload conditions was counterbalanced across subjects.

In the blocked-by-talker group (8 listening for vowel targets, 18 listening for consonant targets), subjects listened for target tokens in spoken sequences of tokens produced by a single talker. Following the set of trials for one talker, the subjects then heard another series of trials with all the tokens produced by a different talker. In this manner, subjects listened to tokens produced by each of the four talkers. The order of the talkers was randomly determined for different experimental sessions under computer control. The other 10 vowel-monitoring subjects and 18 consonant-monitoring subjects listened for target tokens in spoken sequences produced by a mix of four different talkers.

The subjects were given three practice trials in each preload condition to familiarize them with the trial structure and with the recall and monitoring tasks. Following practice, subjects received four blocks of eight trials at the one-number preload and four blocks of eight trials at the three-number preload. Each block consisted of two trials with each of the target vowels /i/, /I/, /u/, and /U/. The sequence of eight trials in each block was randomly determined under computer control. For the blocked-by-talker group, subjects received one block for each of the four talkers in each preload condition; for the mixed-talker group, subjects received the same number of blocks and trials, but the stimuli for each trial were drawn from the set of all four talkers. Thus, subjects in the blocked group heard the same vowel targets and distractors and talkers as the subjects in the mixed group. The only difference was the arrangement of stimuli during trials.

Each trial consisted of a sequence of 16 tokens, each separated by a 250 msec interstimulus interval. The subjects' primary task was to press a response key on a computer keyboard as quickly and as accurately as possible every time a designated target was heard. Four occurrences of a single target were presented at random positions within the sequence of 16 tokens on every trial, with no target presented as the first or last stimulus in a trial, or immediately following a previous occurrence of a target. Each trial began with a short beep produced by the computer and the word READY appearing on the computer screen for three seconds. Following the ready signal, a sequence of randomly selected 2-digit numbers was presented on the Macintosh screen. These numbers were presented one number at a time for two seconds each, with an interstimulus interval of one second. The subjects' secondary task was to remember this sequence of numbers in the correct order throughout the target monitoring trial. The subjects were instructed to perform as accurately as possible on both the monitoring and number-recall tasks.

Next, the target for that trial was displayed on the screen in the form "EE as in beat" for vowels, or "b and in bee" for consonants. After a three second delay, a sequence of stimuli was presented over headphones and subject responses and response times were recorded by the computer. After the 16 tokens in the trial, the computer prompted each subject for the correct sequence of 2-digit numbers that were presented at the beginning of the trial. Finally, the beep and READY sign again appeared signalling the beginning of the next trial.

Results and Discussion

Three measures of target recognition performance were computed for the monitoring task in this experiment: percentage of correct detections (hits), response times (RT) for hits, and percentage of false alarms. Response times were measured from the onset of each stimulus presentation within a trial. Response times under 150 msec were attributed to the immediately preceding stimulus presentation; the response time for that preceding stimulus was computed as the duration of the stimulus plus the interstimulus interval plus the recorded response time. In addition, for each monitoring group, we computed the percentage of numbers correctly recalled in the each of the number-preload tasks. Only those two-digit numbers correctly recalled in their proper position were scored as correct. If a subject correctly recalled a number, but it was in the incorrect position, it was scored as incorrect.

Figure 2.8 shows the mean percentage of 2-digit numbers correctly recalled for the for the one- and three-number preload conditions for both the vowel-monitoring group and the consonant-monitoring group. As would be expected, recall was significantly worse when subjects had to remember three two-digit numbers than when they had to remember a single two-digit number condition, $F(1, 16) = 18.52, p < .0005$ for vowels, $F(1,30) = 26.3, p < .0001$ for consonants. No differences exist in number recall performance between the blocked and mixed groups, nor was the interaction between trial structure (blocked vs. mixed) and preload size significant.

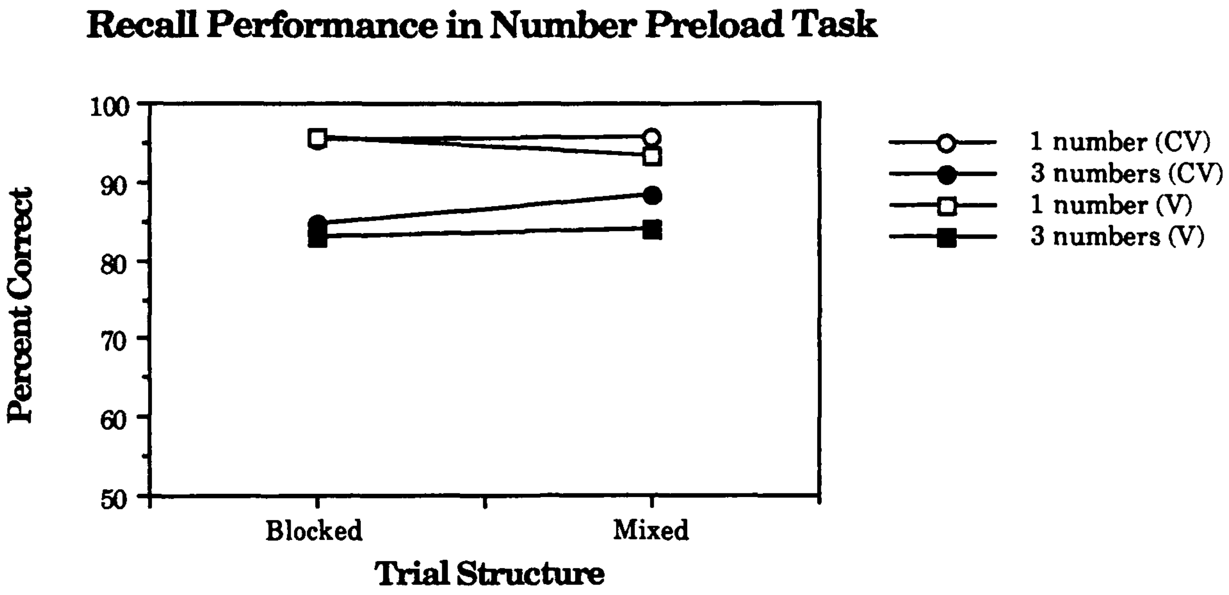

Figure 2.8. Number recall performance in the preload task for consonant and vowel monitoring when each trial consists of speech from one talker at a time (blocked) or a mix of four different talkers.

Figure 2.9 shows the mean correct target recognition rate for the blocked and mixed groups at the two levels of the number preload task (one vs. three two-digit numbers). In general, accuracy was quite good with the hit rate in almost all conditions higher than 90%. Subjects were significantly more accurate with the one-number preload (vowels: 94.3%; consonants: 98.0%) than the three-number preload (vowels: 91.1%, consonants: 95.0%), vowels: $F(1, 16) = 7.56, p < .01$, consonants: $F(1, 30) = 14.8, p<.001$. Thus, increasing cognitive load decreased recognition of vowels and consonants. However, the difference in hit rate between the blocked (vowels: 94%; consonants: 96.4%) and mixed (vowels: 92%; consonants: 96.7%) groups was not statistically significant, nor were any of the interactions significant for hit rate. Thus, target recognition was significantly impaired overall by the increase in central capacity demands made by the greater digit preload condition, although the difference in trial structures (blocked by talker or mixed) did not reliably affect hit rate.

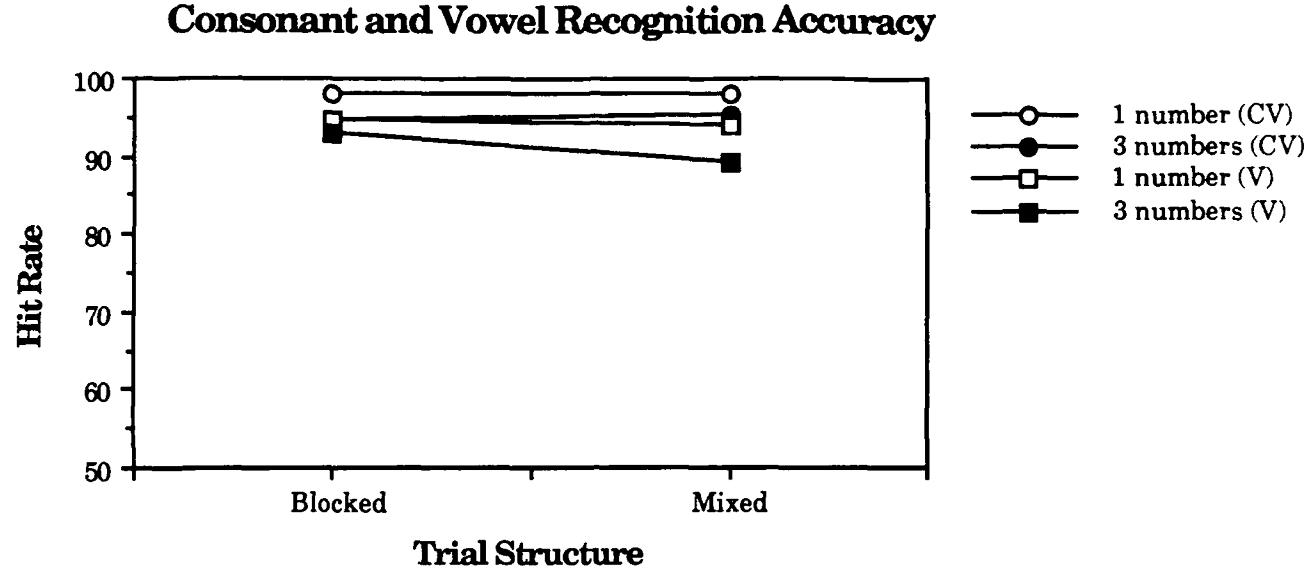

Figure 2.9. Mean correct target recognition in the preload task for consonant and vowel monitoring when each trial consists of speech from one talker at a time (blocked) or a mix of four different talkers (mixed).

Figure 2.10 shows the mean false alarm rates for the four groups of subjects in the one- and three-number preload conditions. Both vowel groups averaged about 2% false alarms in both conditions, and both consonant groups averaged about 1% false alarms. There were no significant effects of the number preload condition or the trials structure (blocked vs. mixed) on the false alarm rates; the interaction between these factors was not significant.

By comparison to the hit rate and false alarm data, the response times for correct responses demonstrate reliable differences between groups and conditions. Figure 2.11 shows the mean response times for the blocked and mixed groups in the two different preload conditions and for the two different token types. Subjects are significantly slower to recognize vowels when the talker varies within a trial than when the talker is constant within a trial, $F(1, 16) = 7.34, p < .02$. When recognizing consonants, however, although subjects were considerably slower in the mixed-talker condition (485.9 ms) than in the blocked-talker condition (442.6 ms), this difference was not significant, $F(1,30) = 2.5$. For the preload conditions, subjects in both the vowel-monitoring and the consonant-monitoring groups are significantly slower to recognize targets when the preload task is more demanding, $F(1, 16) = 6.71, p < .02$ for vowels, $F(1,30) = 6.20, p < .02$ for consonants. The interaction between digit preload and trial structure (blocked vs. mixed), $F(1, 16) = 7.36, p < .02$ for vowels, $F(1, 30) = 5.26, p < .03$ for consonants, can be understood by examining the pattern of response times shown in Figure 2.11. A Newman-Keuls analysis indicates that there is no effect of preload on monitoring times for the blocked group, whereas there is a significant effect of preload on target recognition times for the mixed group ($p < .05$ for both vowels and consonants). Increasing the load on working memory significantly increases response times for recognizing targets when each trial contains a mix of voices. This is a clear demonstration that the slower response times in the mixed condition are due to increased cognitive load imposed by structural estimation.

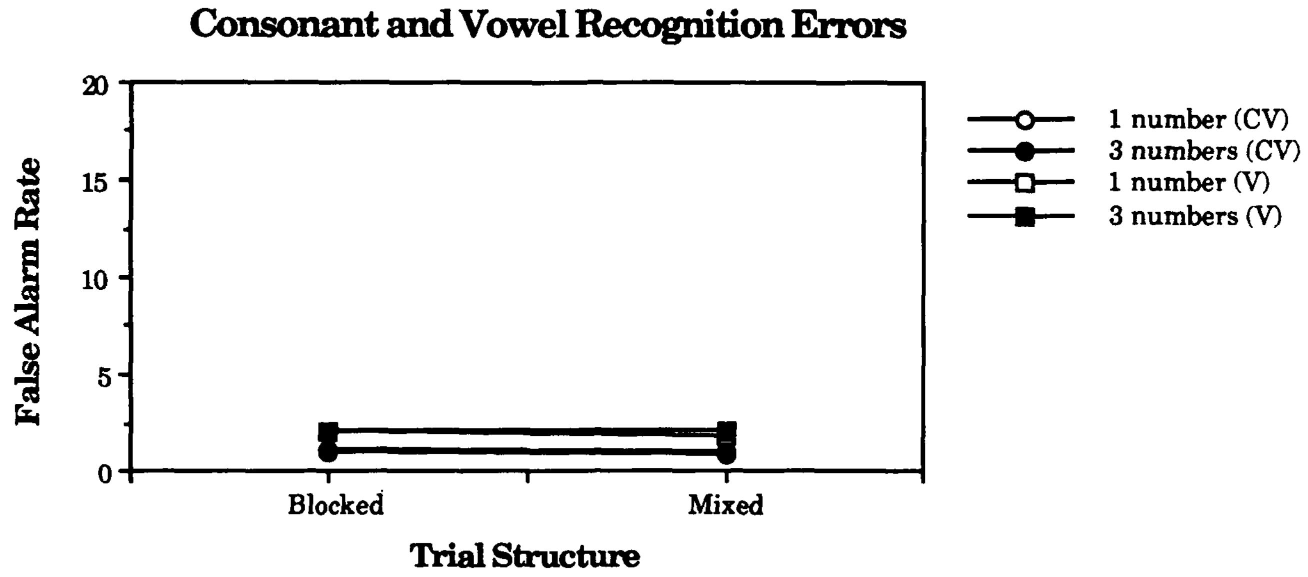

Figure 2.10. False alarms for target monitoring in the preload task for consonant and vowel monitoring when each trial consists of speech from one talker at a time (blocked) or a mix of four different talkers (mixed).

Our earlier experiments demonstrate subjects are slower to recognize vowels, consonants, and words when they hear a mix of talkers compared to a condition in which they listen to a single talker at a time. The present study adds two new pieces of information to our understanding of the perceptual normalization of talker differences in speech perception. First, when listeners use contextual tuning, target recognition does place some capacity demands on working memory indicated by the decreased hit rate for target recognition at the higher preload. This suggests that target recognition per se may be limited by central capacity, as well as by resources specific to speech (Nusbaum & Schwab, 1986). Second, normalization through structural estimation is limited by central capacity. Increasing digit preload places greater demands on working memory which slows target recognition when the talker changes from vowel to vowel. Thus, even though listeners continue to recognize targets across talkers in the mixed condition at a very high level of accuracy, demands on central capacity slow this process even more than when there is no additional load. The operation of this normalization mechanism therefore appears to incur capacity demands that are not made by contextual tuning.

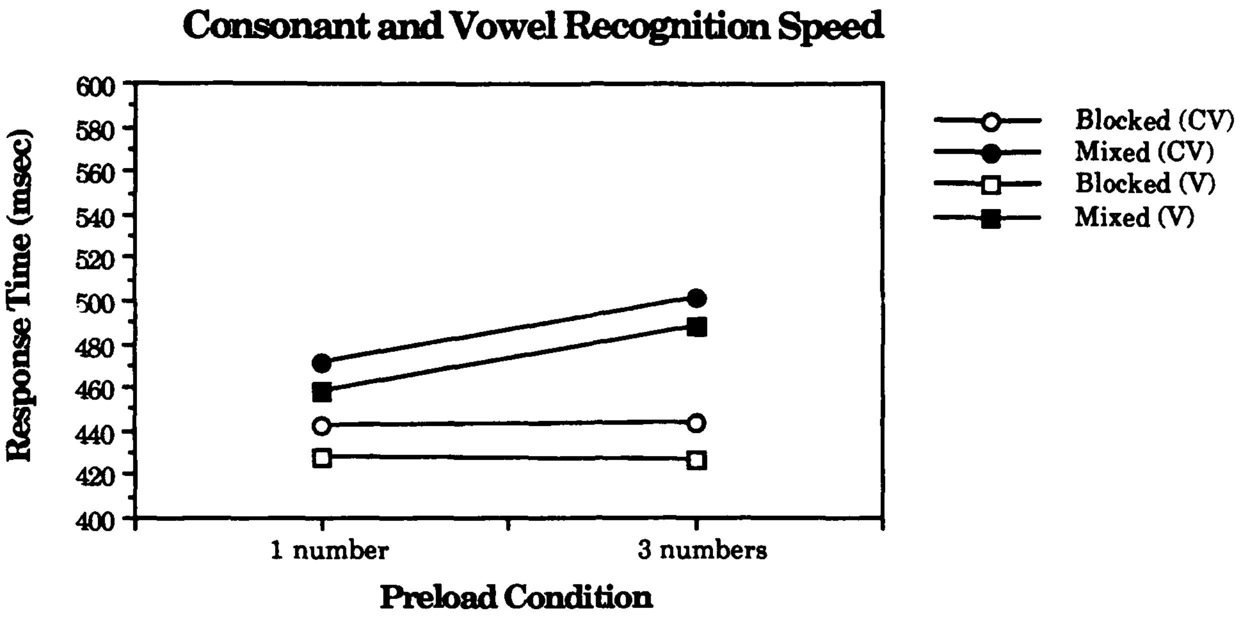

Figure 2.11. Target recognition time for hits in consonant and vowel monitoring when each trial consists of speech from one talker at a time (blocked) or a mix of four different talkers (mixed).

Future Studies

Our results suggest the operation of two different normalization mechanisms in recognition of vowel information in isolation and in CVC contexts, consonant information in CV context, and word information in isolated word context. The contextual tuning mechanism normalizes talker differences based on processing several tokens from the same talker. The structural estimation mechanism normalizes talker differences based on the information contained within a single token, although this requires more effort and attention.

Several questions arise directly from our findings. First, how does the size of a talker's vowel space, or the differences in size between different talker's vowel spaces, figure in the normalization of speech. By varying the number of different talkers and the size of the difference between these talkers in a mixed-talker condition, we can investigate how the magnitude of the differences between talkers affects normalization processes. One group of subjects will hear blocked-talker and mixed-talker conditions with two talkers from the same gender, and the other group will hear two talkers from different genders. Preliminary studies indicate that whereas subjects from both groups are slower in the mixed condition than in the blocked condition, the magnitude of this difference is indeed greater when tokens from two talkers of different genders are mixed than when tokens from two talkers of the same gender are mixed. Thus, structural estimation may require more attention as the vocal characteristics of the mixed talkers becomes more different.

Another question we are investigating concerns the nature of the information that listeners use to normalize speech from different talkers. Syrdal and Gopal (1986) have described a model of structural estimation in which listeners use a talker's pitch and F3 information to recognize vowels across talkers. If we take away these cues (e.g. eliminate pitch information by using whispered vowels) performance should drop accordingly. We have presented subjects with either synthetic reproductions of the vowel stimuli used in Experiment 2.1, or with matched but whispered synthetic vowels missing pitch information. Preliminary studies indicate that subjects are again slower in the mixed-talker condition than in the blocked-talker condition. In addition, however, significantly more errors were made in the mixed-talker than in the blocked-talker condition, and subjects in the whispered vowels groups made reliably more errors than subjects in the synthetic vowels group. Perhaps even more interesting, however, are the accuracy results. Only subjects in the whispered vowels group are significantly less accurate in the mixed-talker condition than in the blocked-talker condition, and by a much larger amount than in our previous studies. To further investigate cues to structural estimation, we will carry out vowel monitoring studies with speech that has been filtered to remove higher formant information. The higher formants are thought to convey information about talker identity. Reductions in vowel monitoring performance for filtered stimuli across blocked and mixed conditions will indicate the differential contribution of this cue to structural estimation.

To investigate the cues to contextual tuning, we will provide listeners with different context sets of vowels and ask subjects to make a speeded recognition judgment about a test stimulus. This context set will serve to provide listeners with the information necessary to calibrate the talker's vowel space. We will examine the influence of the point vowels which delimit the vowel space, nonpoint vowels, and various vowel subsets. Thus, we will vary the size and composition of the target set to determine how this information about a talker's vowel space improves recognition performance based on contextual tuning.

V. Project 3: Effects of Perceptual Learning on Capacity Demands

Synthetic speech is less intelligible than natural speech and recognition of synthetic speech requires more effort and attention than recognition of natural speech (Nusbaum & Pisoni, 1985). In part, these differences in perceptual processing can be attributed to differences in the acoustic-phonetic structures of natural and synthetic speech. Natural speech is rich and redundant in its acoustic-phonetic structure. A variety of different cues may covary across contexts in natural speech to specify phonetic identity. In synthetic speech, by comparison, a much smaller set of acoustic cues is used to convey phonetic information. These cues are typically used to provide minimal contrasts and in some cases, these cues may actually be incorrect or misleading. Thus, synthetic speech is structurally more impoverished than natural speech.

If listeners don't attend to the minimal cue contrasts present in synthetic speech, recognition will be impaired. If listeners do attend to misleading cues, recognition will be impaired. Therefore it is not very surprising that intelligibility is lower for synthetic speech compared to natural speech. Furthermore, recognition of synthetic speech may require more attention because there are fewer cues to attend to and the listener must find and process those cues that are informative. In listening to natural speech, there is so much redundancy that almost any cues will be effective and so less effort is required to find the informative parts of the signal (cf. Nusbaum & Schwab, 1986).

Recently, we have demonstrated that listeners who are given moderate amounts of training with synthetic speech show significant improvements in intelligibility (Greenspan, Nusbaum, & Pisoni, in press; Schwab, Nusbaum, & Pisoni, 1985). In learning to recognize synthetic speech more accurately, listeners may learn which cues are effective in signalling phonetic information. Also, they may learn to reinterpret cues that were previously misleading. This suggests the possibility that as listeners learn to recognize synthetic speech more accurately, the demands of attention for recognition of synthetic speech will be modified following training as attention is more effectively used. Listeners may be learning which aspects of synthetic speech are most informative about phonetic identity, and refocusing their attention on those aspects of the speech.

On the other hand, as a result of training, listeners might actually invest more effort in recognizing synthetic speech. Listeners might learn that recognition of synthetic speech incurs attentional demands and they may simply learn to devote more attention and effort to the recognition process. If this alternative is correct, the effort required for perception of synthetic speech might actually increase following training.

Intelligibility differences between types of speech do affect attentional demands, so that training which improves intelligibility may also affect attentional limitations. Prior research by Luce, Feustel, and Pisoni (1983) has shown that there are differences in recall performance for natural and synthetic speech that suggest greater attentional limitations are imposed by synthetic speech. Preloading short-term memory with different length lists of visually presented digits interacts with the intelligibility of speech in determining recall of spoken words. Also, smaller primacy effects were found for recall of lists of words produced by a text-to-speech system compared to recall of words produced by a human talker.

The present studies were carried out to determine whether training listeners to recognize synthetic speech more accurately affects the attentional limitations of processing the speech. Subjects were given a digit preload task (see Baddeley & Hitch, 1974; Luce et al., 1983) while performing a recall or monitoring task with spoken words prc duced by a text-to-speech system. If attentional demands for perceiving synthetic speech change as a result of training, this should be reflected in an interaction between the demands of digit preload and training in performance on the synthetic speech tasks.

Experiment 3.1: Recall of Synthetic Speech

Luce et al. (1983) demonstrated processing synthetic speech requires more effort and attention than processing natural speech by showing an interaction between type of speech and amount of digit preload. The first study was carried out to determine whether the improvements in intelligibility due to training with synthetic speech would produce attentional effects that mirror the effects produced by intelligibility differences between natural and synthetic speech. If so, we should find an interaction in digit recall between training and size of the digit list. After training, the capacity required for storing synthetic speech should be reduced, thereby increasing the available capacity for digit recall. On the other hand, if training increases effort in processing synthetic speech, the available capacity for digit recall should be reduced by training.

Method

Subjects. The subjects were eight University of Chicago undergraduates, five males and two females, aged 18 to 26 years. All subjects reported English as their native language and had no prior experience with synthetic speech. None of the subjects reported speech or hearing disorders. Each subject was paid $35 upon completion of all five sessions of the experiment. One of the male subjects did not complete the experiment and his data were excluded from analyses.

Stimuli. The materials consisted of 14 lists of 50 phonetically balanced (PB) monosyllabic words (Egan, 1948). The stimuli were produced at a natural-sounding rate by the Votrax Personal Speech System controlled by an IBM-PC/AT microcomputer. The synthetic speech was converted to digital form and stored on an IBM-PC/AT hard disk. Each list of words was sampled through a 12-bit A/D converter at 10 kHz, low-pass filtered at 4.6 kHz as a single waveform file. The waveform file was edited into separate files for each word using a digital waveform editor with 100 microsecond accuracy. The digitized words were converted to analog form at 10 kHz with a 12-bit D/A converter. The speech was low-pass filtered at 4.6 kHz and presented at about 76 dB SPL (measured for the calibration word "cane") over Sennheiser HD-430 headphones.

Procedure. The experiment consisted of five one-hour sessions conducted on each of five consecutive days, Monday through Friday. On the first and last days (Days 1 and 5) of the experiment, the subjects were given a memory task to assess the capacity demands of representing synthetic speech before and after training. On Days 2 through 4 subjects were given training with the Votrax-generated synthetic speech. The memory test and training materials and procedures were identical for all subjects on each day of the experiment. However, during training, subjects never heard the same words twice and different lists of words were used in the pretraining memory test and the posttraining memory test. Experimental sessions were conducted with small groups of 1 to 3 subjects each. An IBM-PC/AT controlled the presentation of stimuli and feedback (when given) and collection of responses from each subject station. Each subject was seated in a sound-attenuating booth in front of a Macintosh Plus computer which presented visual information and collected keypresses.

Pretraining and Posttraining Testing Procedure. Days 1 and 5 were testing days in which subjects were administered a memory task to assess the capacity demands imposed by perception and recall of synthetic speech (Luce, Feustel, & Pisoni, 1983). The memory task was a variation of the digit preload paradigm developed by Baddeley and Hitch (1974). In this task, subjects are given a list of digits to recall prior to another task, such as a list of words to be remembered. In our experiment, a list of either two or four two-digit numbers was presented one at a time on a CRT. Following this list, the subjects heard a list of five words of synthetic speech. The subjects were required first to recall the words they heard and second, to recall the numbers they saw. For both the list of numbers and the list of words, the subjects were asked to recall the items in the order that they were presented. They were encouraged to respond for each of the items presented and to guess for any they were unsure of.

Each testing session was conducted in four blocks of five trials each. Before the first block, a practice trial with a two-number preload and five spoken words was presented. In the first and third blocks a preload of two two-digit numbers was presented before each list of words. In the second and fourth blocks a preload of four two-digit numbers was presented. Thus, in a single testing session, each subject was administered a total of 10 trials with a preload of two two-digit numbers followed by five spoken words and 10 trials with a preload of four two-digit numbers followed by five spoken words.

The words for each block of digit preload testing were sampled with replacement from a single PB list. A different PB list was used for each block. None of the words was ever repeated within a single trial. However, between trials in the same block, the same word could occur more than once. Words for the practice trial were drawn at random from the same PB list used in the first block. Note that although the procedures followed in the digit preload task on Days 1 and 5 were identical, the four PB lists from which the words were drawn were different for each test session.

Training Procedure. During Days 2, 3, and 4 subjects were trained on identification of spoken words produced by the Votrax text-to-speech system. On each training trial, subjects heard a single word and entered their identification of the word onto the computer keyboard. Immediately following the identification response, feedback was provided about the identity of the stimulus. The subjects then simultaneously heard the word spoken again and saw the word printed on the computer screen in front of them. They were then asked whether they had identified the word correctly. After responding Y (Yes) or N (No) to this question, the next training trial began. Subjects were not explicitly told whether they correctly identified the words, nor were they told whether they correctly compared their own identifications with the feedback they received.

Subjects received two blocks of 50 trials on each training day. Each block consisted of 50 words from a single PB list. Although the training procedure was identical for each training session, different pairs of PB lists were used each day. Thus over the course of the three-day training period subjects heard a total of 300 novel stimuli from six PB lists.

Results and Discussion

Subjects typed word responses in both the pretraining and posttraining recall task and during the training session. Word responses were scored as correct if the response correctly matched the target item phonetically, with no missing, permuted, replaced, or added phonemes. For instance, a response to the word "flew" as either "flew" or "flew" would be considered correct. However, responses such as "flute" or "few" or "foo" would be considered incorrect.

In scoring number recall performance, each digit of each two-digit item is considered to be a single unit. Subjects received credit for each single digit recalled accurately and in the correct sequence. For instance, if the sequence " 48 25" were recalled as "48 15" then three digits out of four would have been recalled correctly ("4", "8", and "5"). However, if the sequence were recalled as "45 28" then credit would be given for only two correctly recalled digits ("4" and "2"). A two-digit number recalled correctly but not in the correct sequence relative to the other two-digit numbers was also counted as correct. For instance, if "23 48" were recalled as "48 20", credit would be given for recalling "48" correctly.

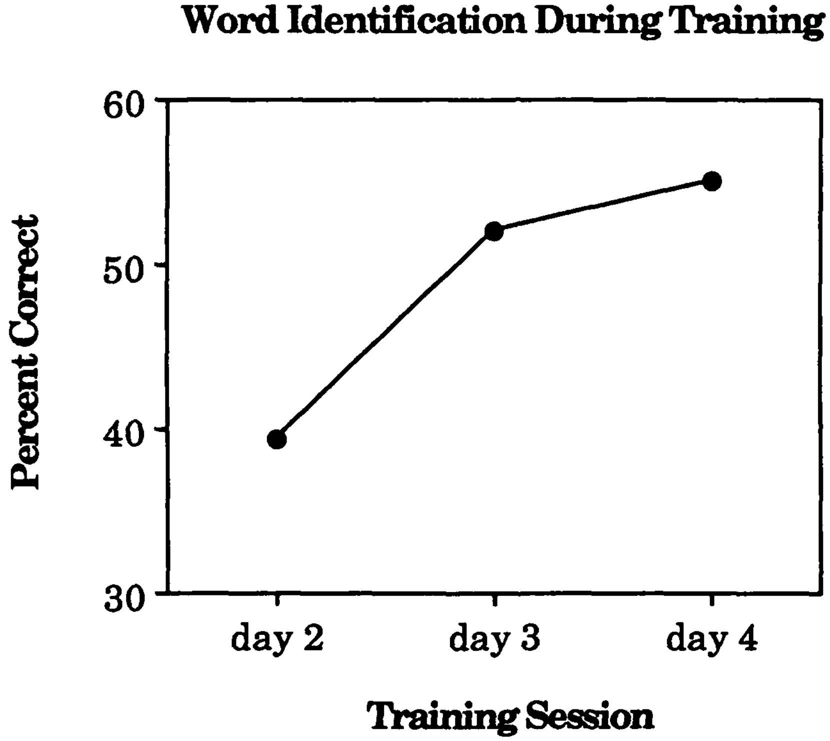

Figure 3.1. Improvements in word identification performance on each day of training with synthetic speech.

Training results. Figure 3.1 shows the improvement in word identification performance as a result of the training procedure. Word identification performance significantly improves from 39.3% correct on Day 2 (the first training session) to 55.1% on Day 4 (the last training session), *F*(2,12) = 39.032, *p* < .01. Post-hoc Newman-Keuls pairwise comparisons revealed that identification performance is significantly better (*p* < .01) on the last two training days than on the first training day, although there was no significant difference in performance between the last two training days. Thus, our subjects significantly improved in identification accuracy for Votrax-generated synthetic speech with a moderate amount of training and perceptual experience replicating our earlier results (Greenspan, Nusbaum, & Pisoni, in press; Schwab, Nusbaum, & Pisoni, 1985). Indeed, performance during training is quite similar to the performance observed in our previous work. It is important to remember that these improvements in word identification performance are all the result of generalization — subjects never heard the same words twice during training.

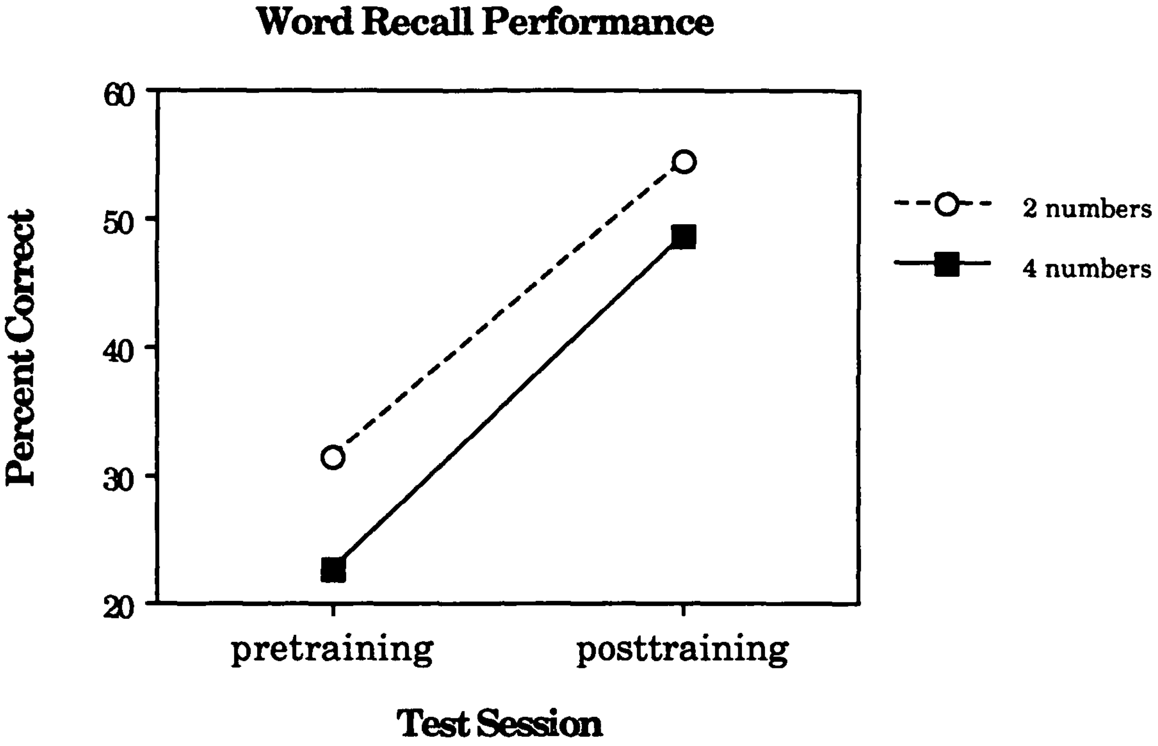

Figure 3.2. Recall of lists of spoken words produced by a text-to-speech system at two levels of digit preload, before and after training.

Word recall results. Word recall performance was assessed before and after training to determine whether changes in intelligibility produced by training modify the capacity demands incurred by synthetic speech (see Luce et al., 1983; Nusbaum, Greenspan, & Pisoni, 1985). Figure 3.2 shows word recall performance at the two levels of digit preload before and after training. As can be seen in this figure, subjects recalled significantly more words after training (51.6% correct) than before training (27.0% correct), $F(1,6) = 68.218, p < .01$. In addition, significantly more words were recalled with two-number (four-digit) preload (43.0% correct), which requires less capacity, than with four-number (eight-digit) preload (35.6% correct), $F(1,6) = 16.127, p < .01$. Thus, training improved recall performance and increasing digit preload reduced recall performance.

Note that there are two ways in which recall performance might improve following training. First, if the capacity demands of synthetic speech are reduced, more synthetic speech could be recalled from short-term storage. Second, if intelligibility increases, word recall performance will also increase for reasons unrelated to capacity. If a word that was presented in a list is identified by the subject incorrectly when it is first heard, it will be scored as incorrect when given as a recall response. As intelligibility increases, the probability of correctly identifying a word increases, thus increasing the probability of correct recall of that word.

In order to determine whether improvements in recall performance following training are due to reductions in capacity demands of synthetic speech or improved word identification, we need to examine the interaction between the digit preload conditions and training. If the difference in recall performance between the two preload conditions decreases after training, we would have evidence for reduced capacity demands as a function of perceptual learning. However, the interaction between the effect of training and digit preload was not significant, *F*(1,6) = .613, *p* > .45. The impact of number preload on word recall did not change significantly as a function of training.

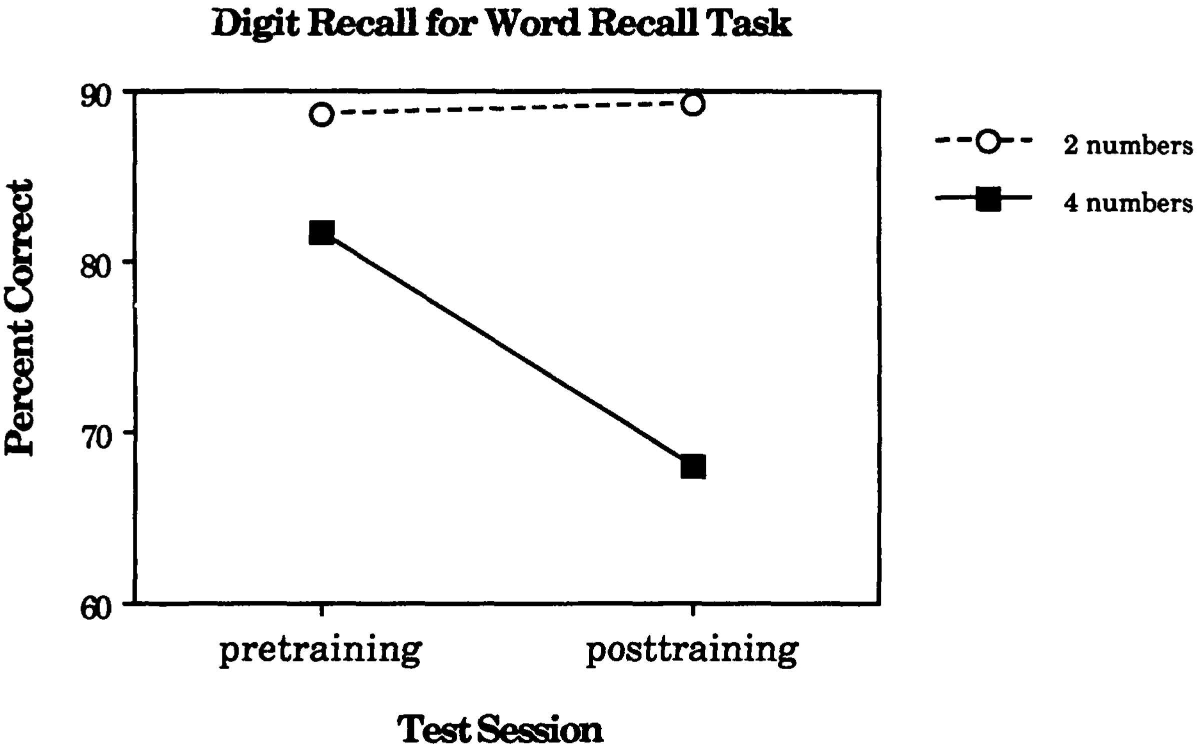

Figure 3.3. Digit recall performance before and after training, for short and long lists of numbers.

Number recall results. Figure 3.3 shows the effects of training on number recall performance for the two- and four-number preload conditions. Following subjects recalled fewer digits than before training, *F*(1,6) = 9.776, *p* < .01. Also, subjects recalled fewer digits with higher memory load (i.e., the longer digit list), *F*(1,6) = 21.549, *p* < .01. Simple effects tests show that there was no effect of training on digit recall at the lowest level of digit preload. However, there was a significant drop in digit recall after training in the four-number preload condition, $F(1,6) = 10.094, p < .05$. Before training, in the four-number preload condition, subjects correctly recalled 81.6% of the digits whereas after training, they correctly recalled only 68.1% of the digits.

Why should number recall performance drop in the four-number preload condition after training? If capacity demands incurred by synthetic speech *decrease* after training, recall performance should improve and not decrease. The answer can be found by considering both the word and number recall data together. Remember that word recall performance increases following training. This may be due in part to increased intelligibility of the words following training. As subjects are able to identify more words correctly, they may be more confident in rehearsing and retaining these items. Thus, following training, subjects may retain more words in memory than they could before training.

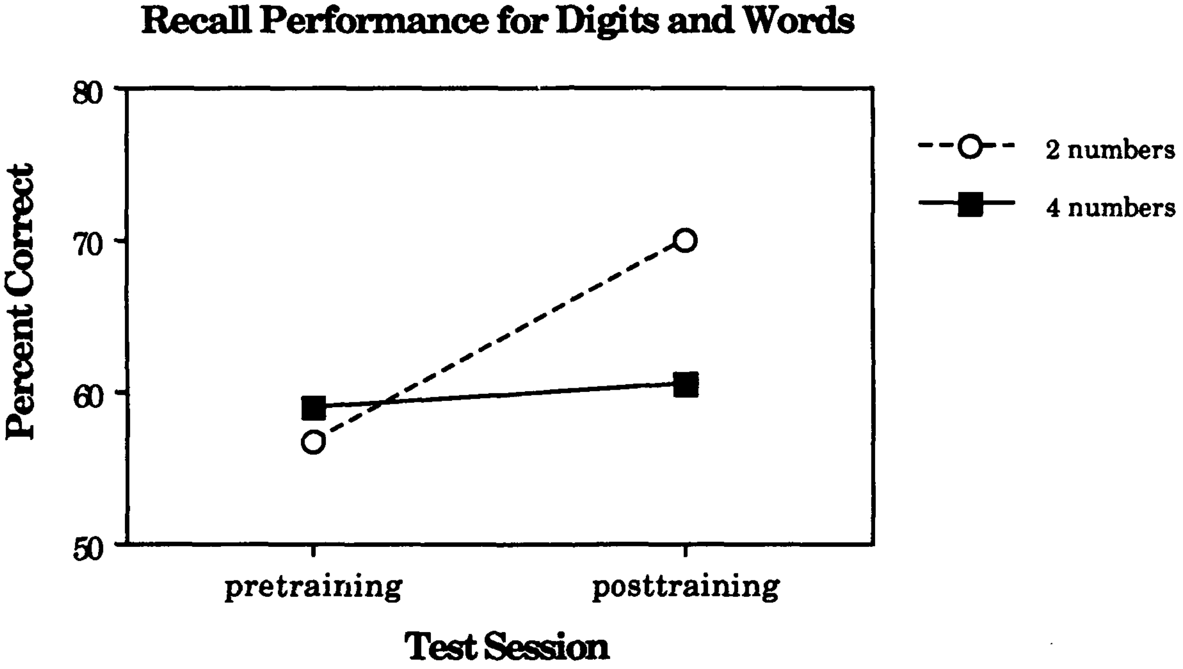

Figure 3.4. Combined recall performance for spoken words and printed digits. At the higher level of digit preload, combined recall is unaffected by training, whereas at the lower level of preload, overall recall improves with training.

Of course, increasing the number of words held in working memory decreases the availability of capacity in short-term memory for digit storage. This suggests that the number of digits recalled should be a function of training and number of to-be-remembered digits, which it is. An examination of the combined recall performance for digits and words together, supports this interpretation, as can be seen in Figure 3.4. For the two-number preload condition, training increases the total number of items recalled. The digit recall data showed no change in the number of digits recalled for the two-number preload between pre- and posttraining memory tests, but word recall increases. The word recall data showed a significant improvement in word recall performance as a result of training. Therefore, the change in total items recalled for the two-number preload condition is due to improved word recall.

The results are somewhat different for the four-number preload condition, also seen in Figure 4. Before training, subjects recalled 58.9% of all items, including digits and words, while after training, subjects recalled 60.5% of all items. Word recall improved significantly, and by the same amount as for the two-number preload condition, but digit recall dropped to compensate for the increased load in working memory. Thus, there is no apparent increase in the available capacity for storing information in memory. Although there are changes in word recall performance due to the intelligibility improvements produced by training, there are no changes in the capacity demands imposed by the synthetic speech.

The results of the present study indicate that perceptual learning of synthetic speech increases intelligibility without affecting the capacity demands imposed by storage and rehearsal of synthetic speech. It may be the case that once synthetic speech is recognized, the capacity demands of storing linguistic representations do not differ whether they are derived before or after training. The differences in the capacity demands observed by Luce et al. may reflect a *data limitation* rather than a *process limitation* (Norman & Bobrow, 1975). In other words, the differences in capacity demands between synthetic and natural speech may be a consequence of the structure of the speech itself, which training does not affect. In order to find effects of training on the processing limitations incurred by perception of synthetic speech, it may be important to examine the real-time recognition processing of synthetic speech before and after training. To investigate changes in the capacity demands due to perceptual encoding, in Experiment 3.2 we carried out a speeded word monitoring task before and after training with synthetic speech.

Experiment 3.2: Speeded Word Recognition

The present experiment was conducted to investigate the effects of training on the capacity demands for recognizing synthetic speech as opposed to storage and retrieval of that speech. Subjects performed a speeded word monitoring task and a digit preload task before and after training. The word monitoring task involved subjects listening to a list of spoken words for a particular target word. Recognition of the target was signalled by pressing a response key as quickly and accurately as possible. In one word monitoring condition, subjects remembered a different list of two two-digit numbers during each monitoring trial. In the other monitoring condition, subjects remembered five two-digit numbers during each monitoring trial. If subjects allocate more attention to perceptual recognition after training, increased digit preload interact with training in slowing responses. If subjects are more efficient in recognizing synthetic speech after training, subjects should be more effective in using spare capacity and digit preload should interact with training in reducing response times.

Method

Subjects. The subjects were 25 University of Chicago students and Hyde Park residents, 11 males and 14 females, aged 17 to 30 years. All subjects reported English as their native language; one subject had prior experience with synthetic speech. None of the subjects reported speech or hearing disorders. Each subject was paid $25 or $30 upon completion of all five sessions of the experiment.

Stimuli. The materials consisted of 14 lists of 50 phonetically balanced (PB) monosyllabic words (Egan, 1948). The speech was produced, digitized, edited and presented as individual stimuli as described in Experiment 3.1.

Procedure. The design was similar to that of Experiment 3.1. The experiment consisted of five one-hour sessions conducted on each of five consecutive days (Monday through Friday), or on a single day for the pretraining test session, plus four consecutive days for training and the posttraining test (Friday, and Monday through Thursday). On the first and last days of the experiment (Days 1 and 5), the subjects were given a speeded word monitoring task with digit preload to assess the capacity demands of recognizing synthetic speech before and after training. On Days 2 through 4 they were given training with the Votrax-generated synthetic speech. The training materials and procedures used for Days 2 through 4 were identical for all subjects. The materials for the monitoring task were also identical; however, the order of presentation for the preload conditions and the PB lists was counterbalanced.

Pretraining and Posttraining Testing Procedure. The word monitoring task was administered on Days 1 and 5 using a variation of the digit preload paradigm developed by Baddeley and Hitch (1974). Subjects were given a list of digits to recall prior to performing a speeded word recognition task. Subjects first saw a list of either two or five two-digit numbers, presented one at a time on a CRT. After seeing the list of numbers, subjects heard a list of 16 synthetically produced spoken words separated by a 250 msec ISI. They monitored this list for a designated target word that was presented on the CRT prior to the presentation of the word list. After the monitoring task, the computer prompted subjects to recall the numbers they saw on the screen, in the order they were presented. All subjects began each testing session with a short practice block in which they monitored for a target word in a series of spoken words, but they did not see numbers for later recall. The subjects were then administered two test blocks, one in which they received a digit preload of two numbers and another in which they received a digit preload of five numbers. The order of the preload conditions was counterbalanced across subjects.

Before each trial began the computer produced a short beep to alert subjects to the beginning of the trial. The word READY then appeared on the screen and remained there for three seconds. After the ready signal, the computer presented a series of randomly selected two-digit numbers. These numbers appeared on the screen one at a time for two seconds each, with an interstimulus interval of one second. After presentation of the digit preload, the target word appeared on the screen. The target word remained on the screen throughout presentation of each series of words. Three seconds after the target word appeared on the screen the computer presented a series of 16 words chosen from a single PB list. The target

was presented four times at random locations throughout the series of words, with the exception that targets never appeared first or last in a series, and they never appeared in consecutive positions. The computer selected a different target word for each trial in a block. The target word for one trial could be randomly selected by the computer as a filler item for any of the other trials in the block. In addition, the computer could randomly select the same filler item to appear more than once, and could present it in consecutive positions, in the same trial.
A total of five PB lists was used during Days 1 and 5 combined. The words for a single block were sampled from a single PB list. For the practice blocks on both Days 1 and 5, words were drawn from one of the five PB lists. Two different PB lists were used for the two test blocks on each of these days. All subjects heard one set of two PB lists during the test blocks on Day 1, and the other set of two PB lists during the test blocks on Day 5. The order of presentation for the two lists on each day was counterbalanced.

Training Procedure. During Days 2, 3, and 4 subjects were trained on identification of spoken words produced by the Votrax text-to-speech system. The procedure for each training trial in the present was identical to that in Experiment 3.1. Subjects received three blocks of 50 trials on each training day. Each block consisted of 50 words from a single PB list. Although the training procedure was identical for each training session, different sets of three PB lists were used each day, so that subjects never heard the same words twice during training. Thus over the course of the three-day training period subjects heard a total of 450 novel stimuli from nine PB lists.

Results and Discussion

Scoring procedures for word identification in the training sessions and for digit recall in the digit preload task were the same as outlined for Experiment 3.1. For the word monitoring task, we computed three measures of word recognition performance: percentage of correct detections (hits), response times (RT) for hits, and percentage of false alarms. RTs were measured from the onset of each stimulus presentation in each trial. RTs of less than 150 ms were assigned to the immediately preceding stimulus in the series. The RT for this preceding stimulus was computed as the duration of the stimulus plus the interstimulus interval plus the recorded response time of less than 150 ms recorded for the following stimulus.

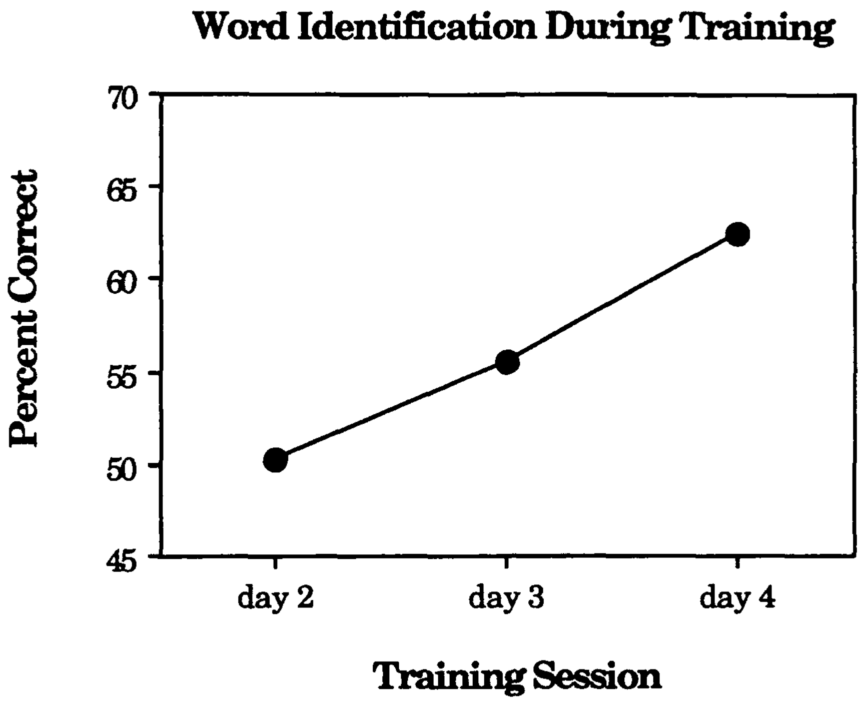

Figure 3.5. Effects of training on word identification accuracy.

Training Results. The training procedure was highly effective in increasing subjects' ability to recognize synthetically produced speech. Figure 3.5 shows that subjects improved in word identification performance as a result of the training procedure, *F*(2,48) = 94.448, p < .01. Their performance improves from 50.24% correct identification on Day 2 (the first training session), to 55.47% on Day 3, to 62.45% correct identification on Day 4 (the last training session). Post-hoc Newman-Keuls pairwise comparisons indicated that performance improves significantly between the first and second training days (*p* < .01), and between the second and third training days (*p* < .01).

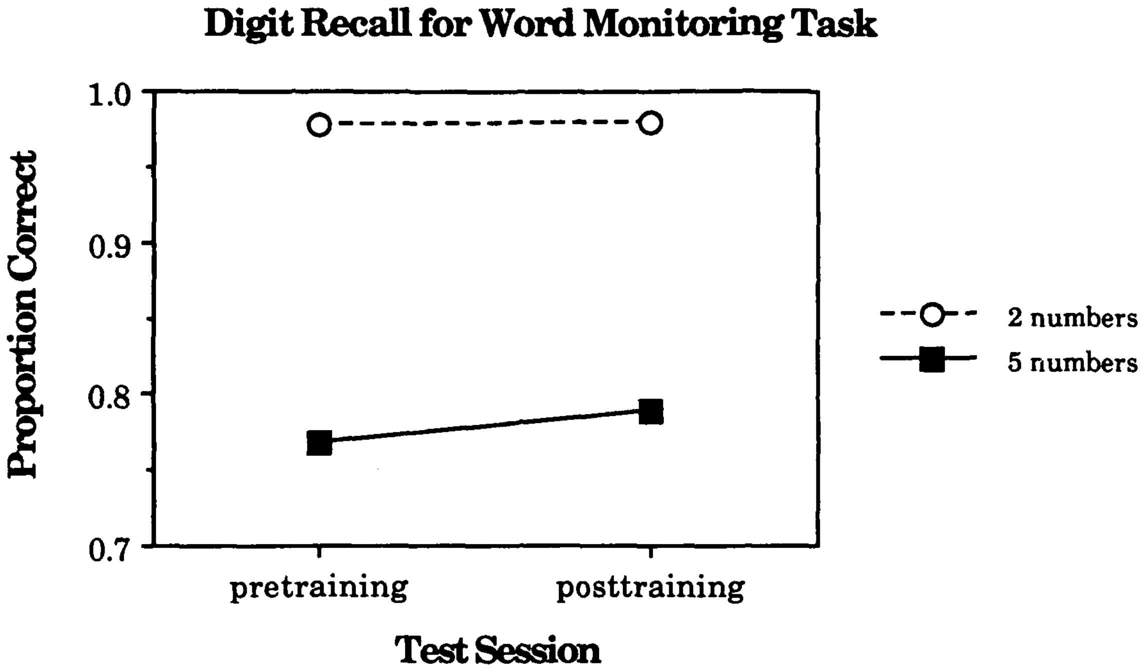

Figure 3.6. Digit recall performance before and after training for short and long lists of two-digit numbers.

Number Recall Results. Training did not affect digit recall performance on the digit preload task (see Figure 3.6). The only factor that affected digit recall was preload condition. Mean accuracy was higher in the low preload condition (97.84%) than in the high preload condition (77.78%), *F*(1,24) = 102.169, p< .01. The interaction between training and preload was not significant. Thus, subjects had less capacity available for the word monitoring task in the high preload condition than in the low preload condition both before and after training.

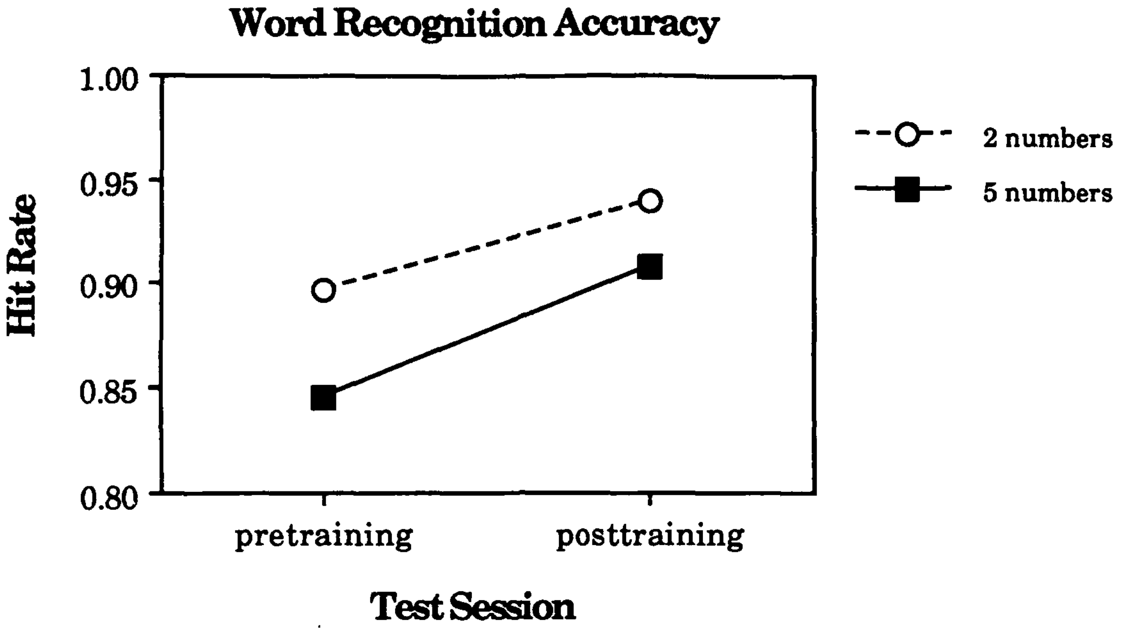

Figure 3.7. Word recognition accuracy in the monitoring task before and after training, with short and long lists of numbers for digit preload.

Word Monitoring Results. Decreasing available capacity through the digit preload task reduced subjects' ability to recognize synthetic speech, whereas training increased the ability to recognize synthetic speech. Figure 3.7 shows the effect of training and preload condition on word recognition accuracy (hit rate) in the monitoring task. Subjects showed significantly higher hit rates in the low preload condition (91.82%) than in the high preload condition (87.70%), $F(1,24) = 15.677$, $p < .01$, and showed significantly higher hit rates after training (92.38%) than before training (87.14%), $F(1,24) = 15.622$, $p < .01$. The interaction between training and preload for hit rate was not significant.

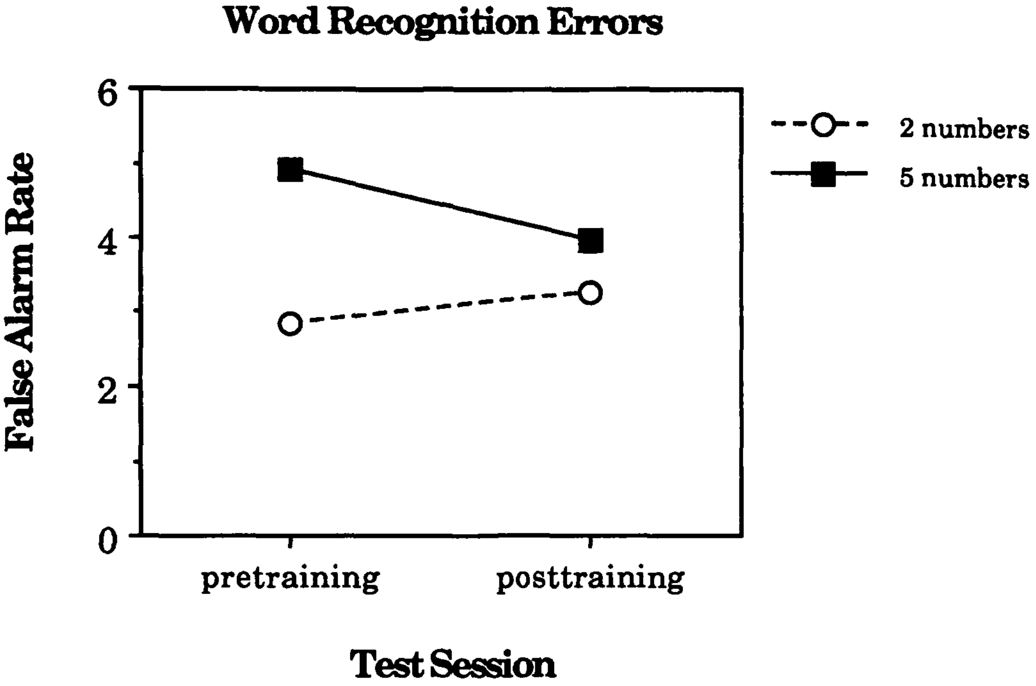

Figure 3.8. The mean number of false alarms in the word monitoring task before and after training, with low and high levels of cognitive load based on the digit preload task.

Subjects made significantly more word monitoring errors when they had less capacity available for the word recognition. Figure 3.8 shows the mean number of false alarms in the low and high preload conditions before and after training. Subjects showed a significantly higher number of false alarms in the high preload condition (4.44) than in the low preload condition (3.06), *F*(1,24) = 10.665, *p*< .01. However, the number of false alarms was not significantly different before training (3.88) and after training (3.62), nor was the interaction between training and preload significant.

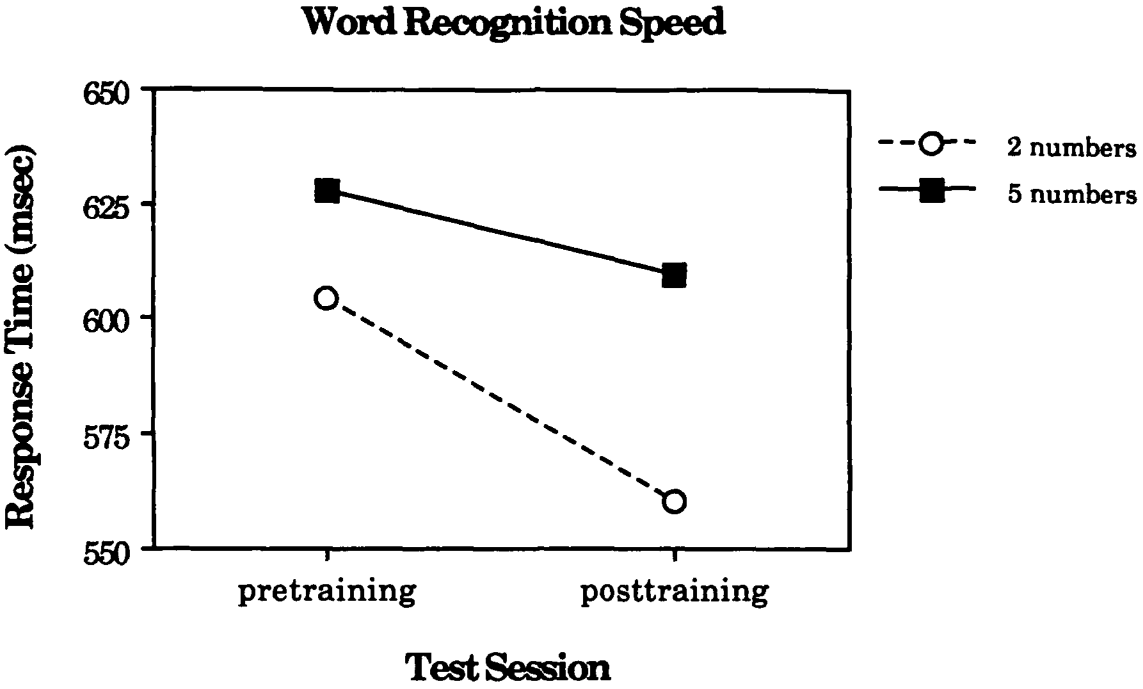

Figure 3.9. Mean word recognition times for synthetic speech before and after training, at the low and high levels of cognitive load in the digit preload task.

Figure 3.9 shows the mean response times in the low and high preload conditions before and after training. Subjects recognize significantly faster in the low preload condition (582.2 ms) than in the high preload condition (618.7 ms), *F*(1,24) = 22.300, *p*<.01, and they recognize synthetic speech faster after training (584.9 ms) than before training (616.0 ms), *F*(1,24) = 5.802, *p*< .05. The interaction between training and preload was also significant *F*(1,24) = 4.453, p < .05. At the lower preload, when subjects had more spare cognitive capacity, training reduced word recognition time even more than at the high preload, when there was little excess capacity available. Simple effects tests showed that both before and after training, RTs were significantly faster i.: the low preload condition than in the high preload condition, *F*(1,24) = 5.204, *p* < .05 and *F*(1,24) =30.242, *p*< .01, respectively. For the low preload condition, RTs were significantly faster after training, *F*(1,24) = 9.94, *p* <.01; however, for the high preload condition RTs were not significantly different before and after training.

These results suggest that listeners may be learning to refocus their attention to the more informative parts of the acoustic-phonetic structure of synthetic speech. By directing attention in a way that is appropriate to synthetic speech, listeners can use space cognitive capacity more effectively. Learning to recognize synthetic speech may be a process of learning to direct phonetic or auditory attention to compensate for acoustic-phonetic differences between natural and synthetic speech. Perhaps an inappropriate focusing of attention creates greater attentional demands for processing synthetic speech, which results in lower recognition ability.

An alternative explanation for our results, however, may be postulated. Perhaps the increased intelligibility of synthetic speech causes the differences in attentional limitations before and after training. Recognizing synthetic speech requires more attention than natural speech and synthetic speech is less intelligible that natural speech. Similarly, before training, recognizing synthetic speech requires more attention and is less intelligible than after training. On the one hand, training may focus attention, thus reducing demands and increasing intelligibility. On the other hand, training may increase intelligibility thus decreasing attentional demands. It may be the improvements in the ability of the listener to recognize synthetic speech that affects attentional limitations, rather than the refocusing of attention that improves recognition. In Experiment 3.3 we improved subjects' ability to recognize synthetic speech not through training, as in the first two experiments, but by using higher quality synthetic speech. If it is simply the intelligibility of synthetic speech that affects attentional demands, then if we raise the intelligibility of the speech without training the listener, we should see an interaction in recognition speed between intelligibility of the speech and digit preload similar to the interaction between training and digit preload.

Experiment 3.3: Effects of Intelligibility on Attentional Limitations

In this experiment we varied the listeners' ability to recognize synthetic speech by using speech from two different text-to-speech systems, the Votrax Personal Speech System and the CallText 5000 System. Votrax speech is much less intelligible than CallText speech (Greene, Logan, & Pisoni, 1986). Two groups of subjects once again performed the digit preload and word monitoring tasks, as did the subjects in Experiment 3.1, but they received no training. One group of subjects heard the less intelligible Votrax speech, and another group heard the more intelligible CallText speech.

Method

Subjects. The subjects were 47 University of Chicago students and Hyde Park residents, 26 males and 21 females, aged 17 to 33 years. All subjects reported English as their native language and none of the subjects reported speech or hearing disorders. Each subject was paid $4 upon completion of the experiment. The data from five subjects were excluded from the analyses. Of these five, three were excluded because of of equipment failure, one because her native language was not English, and one because he was unable to perform the task.

Stimuli. The materials consisted of five lists of 50 phonetically balanced (PB) monosyllabic words (Egan, 1948). One set of lists were produced by the Votrax Personal Speech System and a second set were produced by the CallText 5000 text-to-speech system. The waveforms were digitized, edited and presented as stimuli as described in Experiment 3.1.

Procedure. The experiment was carried out in a single one-hour session. Each subject participated in the same digit preload and word monitoring tasks as described in Experiment 3.2. The subjects were assigned to one of two experimental groups. The first group listened to speech produced by the less intelligible system, the Votrax system. The second group listened to more intelligible speech produced by the Calltext system.

A total of three PB lists was used during one experimental session for one group of subjects. Words from the same PB list were used in the practice block for all subjects. In the test blocks subjects heard words from two of the remaining four lists. The same two lists were presented to about half of the subjects in each condition. The order in which the two lists were presented was counterbalanced. The remaining two lists were presented to the other half of the subjects in each condition, with the order of list presentation counterbalanced.

Results and Discussion

Scoring procedures for the training sessions and the digit preload task were the same as outlined for Experiment 3.1. Scoring procedures for the word monitoring task were the same as outlined for Experiment 3.2.

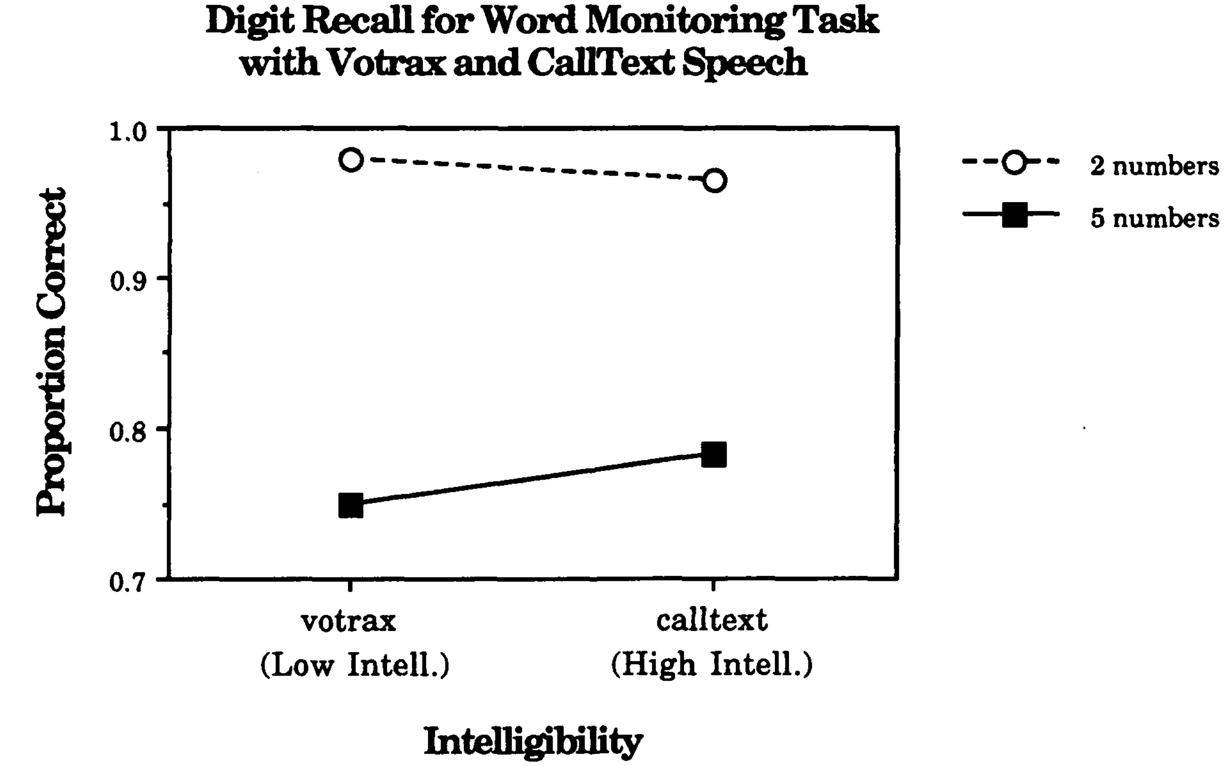

Figure 3.10. Recall performance for short and long lists of visually presented digits for subjects who listened to Votrax- and CallText-generated speech.

Number Recall Results. Figure 3.10 shows the effect of type of intelligibility (Votrax vs. CallText speech) on number recall performance for the two-number (low) and five-number (high) preload conditions. Digit recall was better at the lower preload level (97.26%) than in the high preload condition (76.60%), *F*(1,40) = 112.140, *p*< .01. However, mean digit recall performance was significantly different for Votrax (86.45%) and Calltext speech (87.40%), nor was there a significant interaction between training and preload.

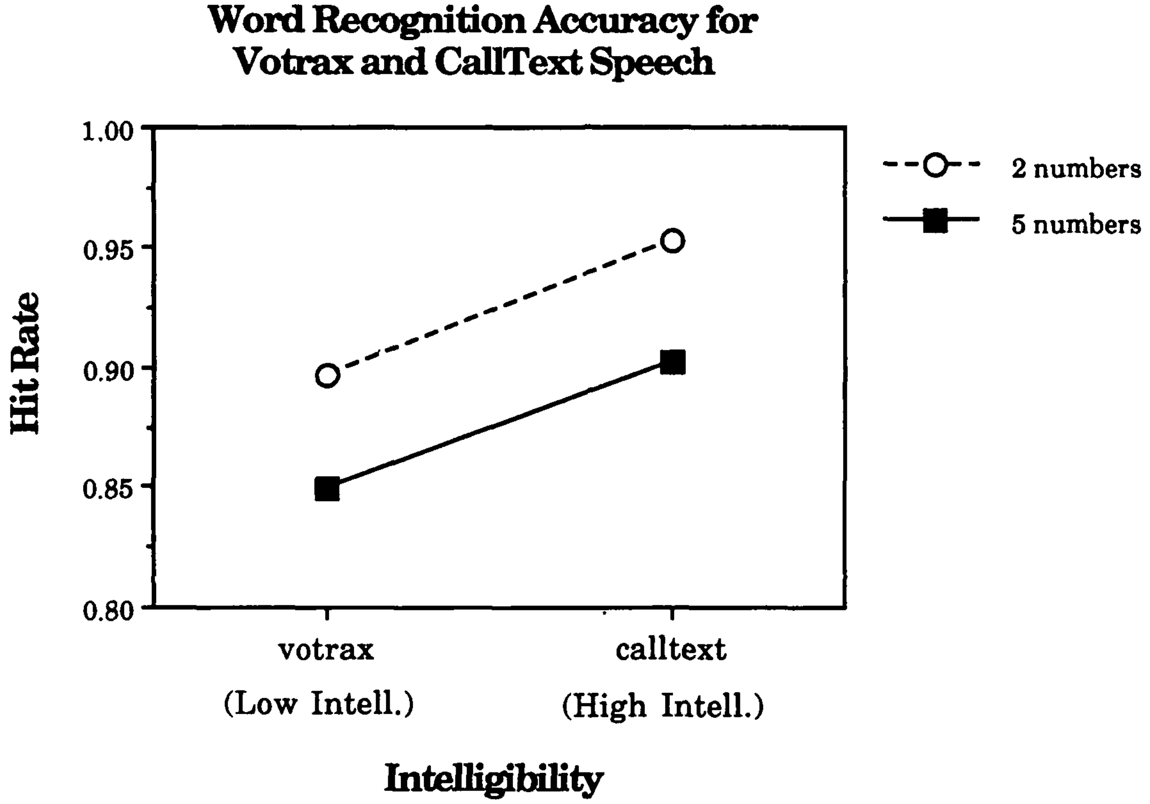

Figure 3.11. Word recognition accuracy for Votrax- and CallText-generated synthetic speech under conditions of low- and high-levels of digit preload.

Word Monitoring Results. Figure 3.11 shows the effects of speech and preload condition on word recognition accuracy (hit rate) in the monitoring task. Subjects showed significantly higher hit rates in the low preload condition (92.43%) than in the high preload condition (87.60%), *F*(1,40) = 13.337, *p* < .01, and showed significantly higher hit rates with Calltext (92.74%) than with Votrax speech (87.29%), *F*(1,40) = 6.114, *p* < .05. Thus, recognition accuracy was higher for the more intelligible speech and increasing cognitive load reduced recognition accuracy. The interaction between type of speech and preload condition for hit rate was not significant.

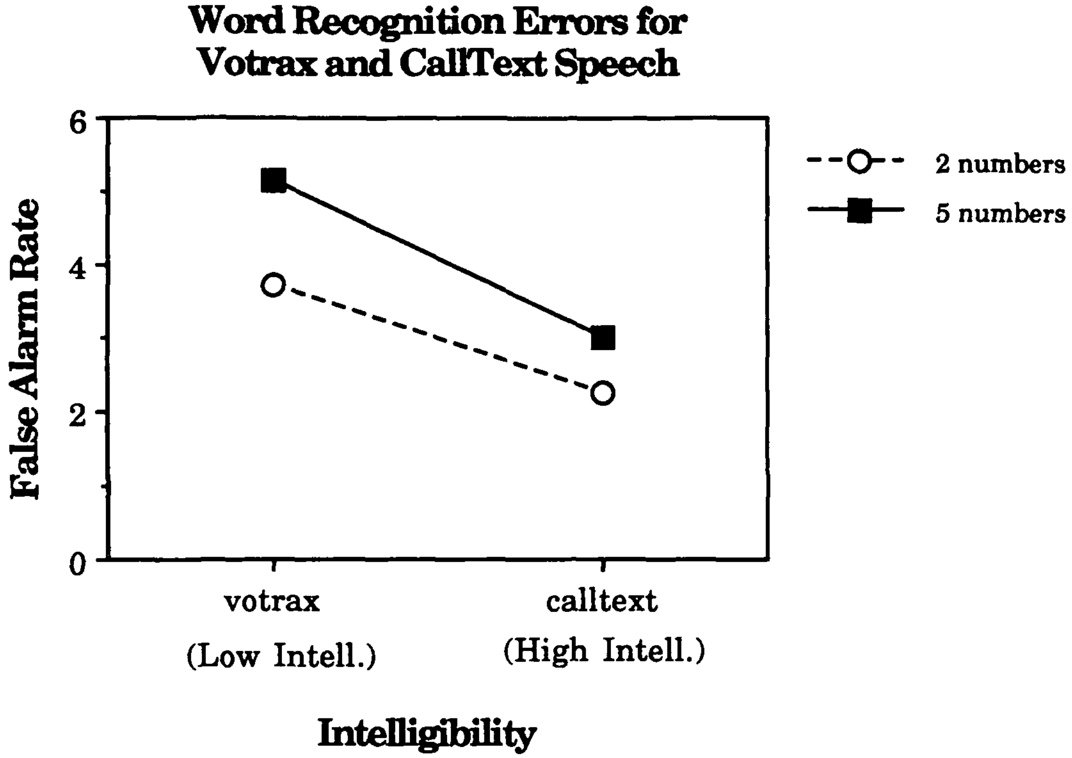

Figure 3.12. False alarms in the monitoring task for Votrax- and CallText-generated synthetic speech under conditions of low and high cognitive load in the digit preload task.

Subjects made fewer word recognition errors with the more intelligible synthetic speech, but there was no effect of digit preload on error rates. Figure 3.12 shows the mean number of false alarms in the low and high preload conditions with each type of speech. Subjects showed a significantly higher mean number of false alarms for Votrax speech (4.42) than for Calltext (2.62), $F(1,40) = 7.215, p < .01$. The difference in mean number of false alarms for the low (2.98) and high (4.07) preload conditions was not significant, nor was the interaction between type of speech and preload significant.

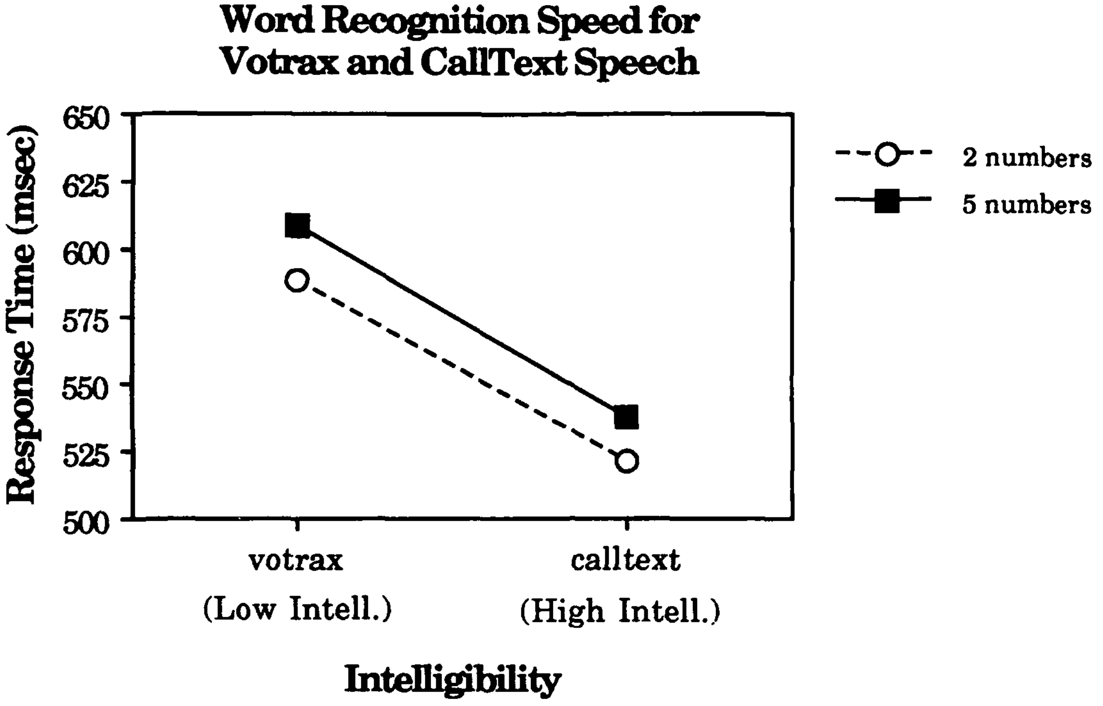

Figure 3.13. Word monitoring times for Votrax- and CallText-generated synthetic speech at low- and high-levels of cognitive load in the digit preload task.

Figure 3.13 shows the mean response times in the low and high preload conditions for each type of speech. Word recognition was significantly faster in the low preload condition (554.8) than in the high preload condition (573.0), $F(1,40) = 6.348$, $p < .05$. Response times were also faster for words produced by the CallText system (529.6) than for words produced by the Votrax system (598.2), $F(1,40) = 6.905$, $p < .0121$. Remember that one interpretation of our results in the previous study was that increasing intelligibility (as a result of training or by any method) should improve the effective use of cognitive capacity. If this were true, the pattern of response times in the present study should mirror the interaction we observed between training and cognitive load. However, in the present study, the intelligibility of the synthetic speech did not interact with cognitive load. It is important to note that our manipulation of intelligibility by using a different text-to-speech system in the present study matches closely the effects of training: A comparison of the hit rates and response times before and after training on Votrax speech with the hit rates and response times in the present study demonstrates a fairly close correspondence. Thus, the intelligibility of the CallText speech was about the same as Votrax speech after training. The absence of a significant interaction between preload and type of speech suggests that simply making synthetic speech more intelligible does not result in more efficient processing of the speech. Rather, it is perceptual learning that enables subjects to use spare cognitive capacity more efficiently when processing synthetic speech perhaps by focusing attention on the relevant acoustic-phonetic properties of the speech.

Future Studies

We have begun to develop a model of the interactions between learning of synthetic speech and the changes in attentional focus. This model will serve as the basis for making predictions in future experiments and represents an explanation of the way perceptual learning may increase the efficient use of cognitive resources by appropriately directing the focus of attention during speech perception. In addition, this model is based on a set of theoretical principles that may generalize to account for other phenomena relating attention and learning in speech perception, such as our research on perceptual normalization of talker differences.

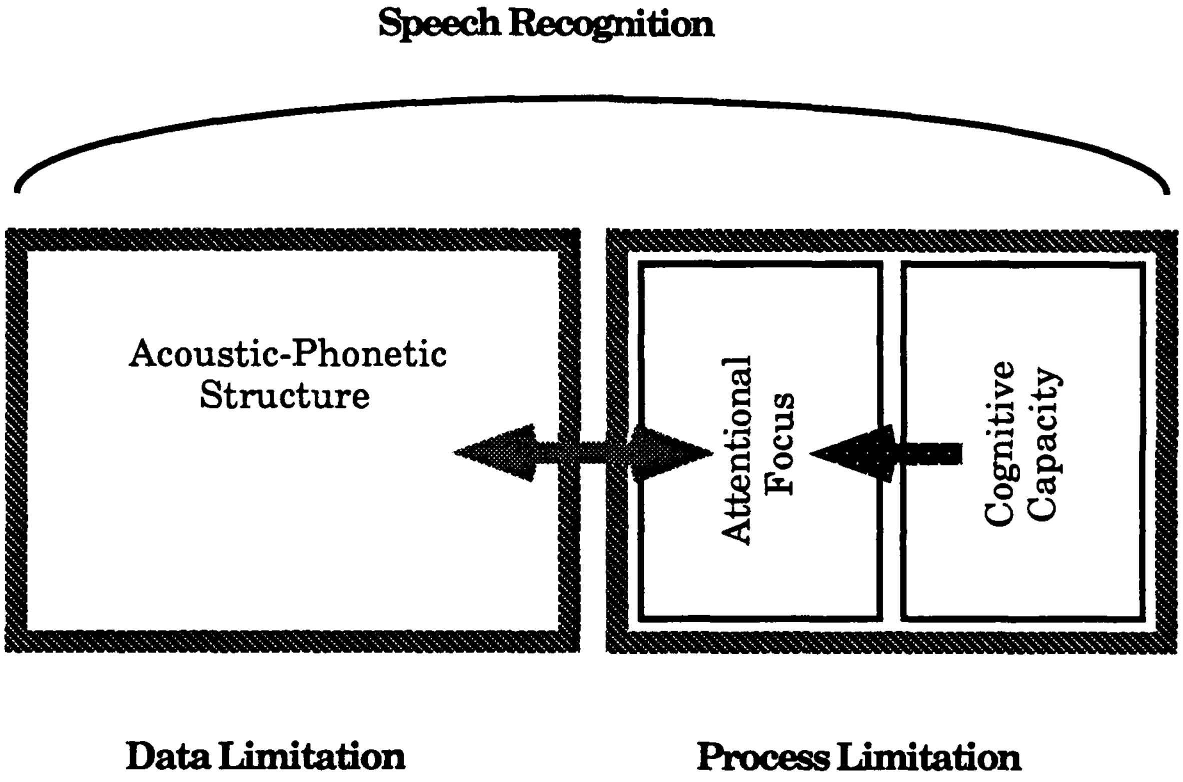

Figure 3.14. Framework for a model relating perceptual learning to attentional limitations in speech perception. Available cognitive capacity is directly through attentional focus to different parts of the acoustic-phonetic structure of speech. Note that this figure is not intended to represent a flowchart, but instead identifies the relations among the major theoretical constructs in our model.

The overview of the model, shown in Figure 3.14, divides performance limitations in speech perception into two components, following the terminology of Norman and Bobrow (1975). Data limitations constrain performance based on the information present in the acoustic-phonetic and prosodic structure of the speech. For synthetic speech, these limitations are severe by comparison with natural speech, because synthetic speech incorporates fewer of the cues to phonetic distinctions than natural speech. However, there are limitations on the processing of this structure by the listener. Process limitations are composed of the available cognitive capacity of the listener and the allocation or focus of that capacity on the acoustic properties of the speech.

The intelligibility of speech is a function of the acoustic-phonetic structure, available cognitive capacity, and the use of that capacity. Changing cognitive load by the digit preload task modifies the amount of cognitive capacity that can be used in recognizing spoken language. Different text-to-speech systems produce speech with different acoustic-phonetic structure thereby affecting the data limitations on listener performance. Perceptual learning allows the listener to focus capacity more effectively on the acoustic-phonetic structure of speech, thereby raising intelligibility *and* increasing the rate of information transfer from the speech signal.

Using these components we have simulated performance in the speeded word monitoring task under different digit preloads. We start by assuming that the listener allocates some capacity to the digit preload task before word monitoring begins, and then allocates the remaining capacity to the word recognition task. The allocation strategy for the digit preload task is to take increasingly smaller proportions of cognitive capacity for each succeeding number to be remembered. This conservative strategy ensures that there will be sufficient capacity for word monitoring, and has a side effect of explaining a finding that has puzzled Baddeley (1986) for some time.

Baddeley (1986) suggested that if working memory were a central limitation on performance of two tasks, such as digit preload and word recall or monitoring, there should be an inverse correlation between performance on the two tasks. As performance on one task improves due to increased capacity allocation, performance on the other task should decrease due to the loss of resources. However, Baddeley has generally found that when digit preload increases performance drops on *both* the digit task and the other task (e.g., word monitoring), which is also what we observed in our studies.

The capacity allocation strategy outlined above produces this effect. Increasing digit preload takes capacity away from word monitoring, but each successive digit does not take as much capacity as it might need for perfect recall performance. In fact the amount of capacity allocated to digit recall decreases with each digit so that performance drops on both tasks or increases on both tasks with the amount of digit preload.

The next part of the model assumes that the recognition of a word follows a time-accuracy growth function that we simulate by an inverse exponential. The asymptote is the ultimate intelligibility of the speech determined by the product of the data limitation and process limitation factors. The rate of approach to the asymptote is a function of the available capacity and the focus of attention. When attention focuses capacity in a way that is appropriate to the acoustic-phonetic structure of the signal, the slope of the gr^wth function increases meaning that higher accuracies are achieved in shorter durations. Figure 3.15 shows the time-accuracy curves generated by the simulation.

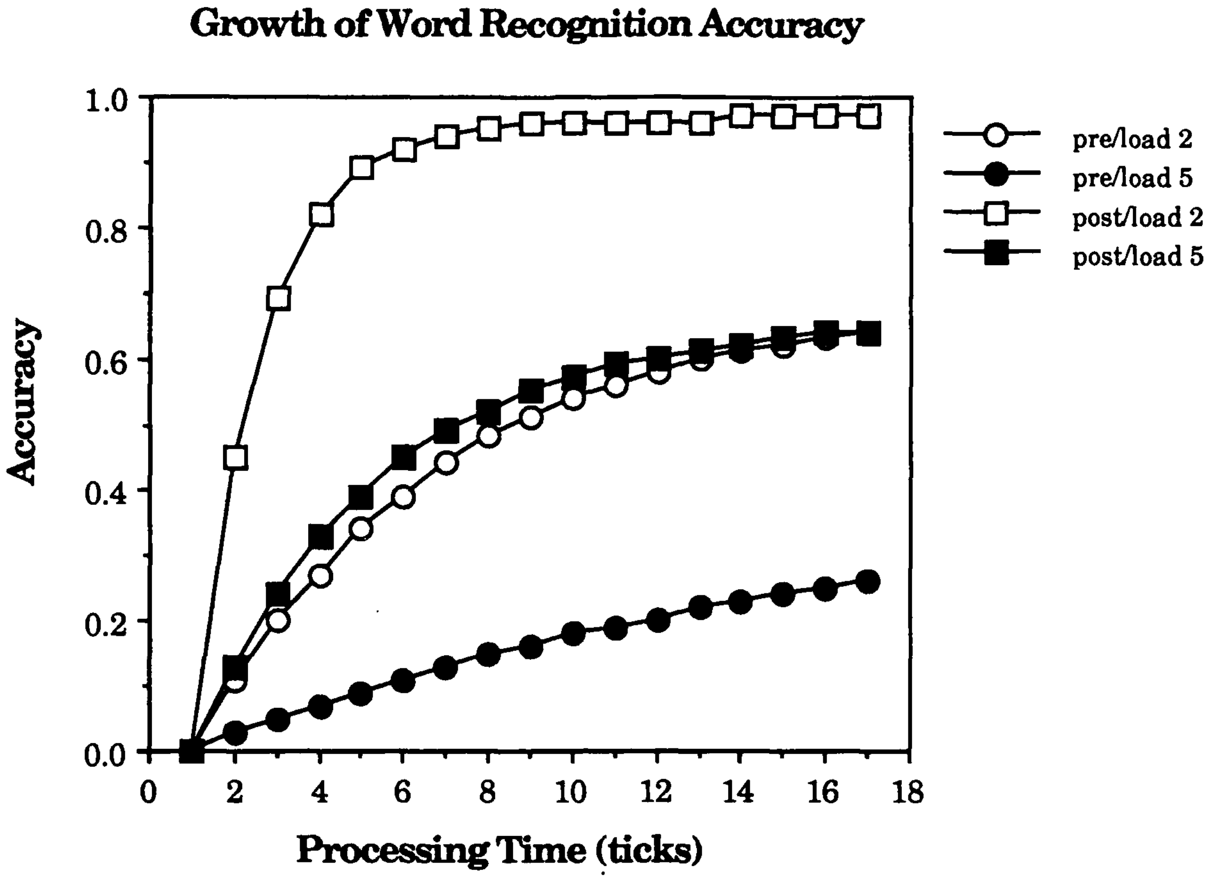

Figure 3.15. Effects of perceptual learning and cognitive load on the probability of a correct word recognition response. Performance on pretraining tasks is shown by circles and after training by squares. Performance on word recognition when the demands of the digit preload task are small is shown in the open symbols and performance at the high level of digit preload is show by the filled symbols.

Adopting a particular accuracy criterion allows us to derive simulated hit rate and response time data from the growth curves in Figure 3.15. The hit rate data produced by the model are shown in Figure 3.16 and the response time data are shown in Figure 3.17. Training and increasing available capacity produce an interaction in response times but not accuracy levels. This is the pattern of results we observed in our second experiment in which subjects performed a word monitoring task before and after training at low and high levels of cognitive load specified by the digit preload task.

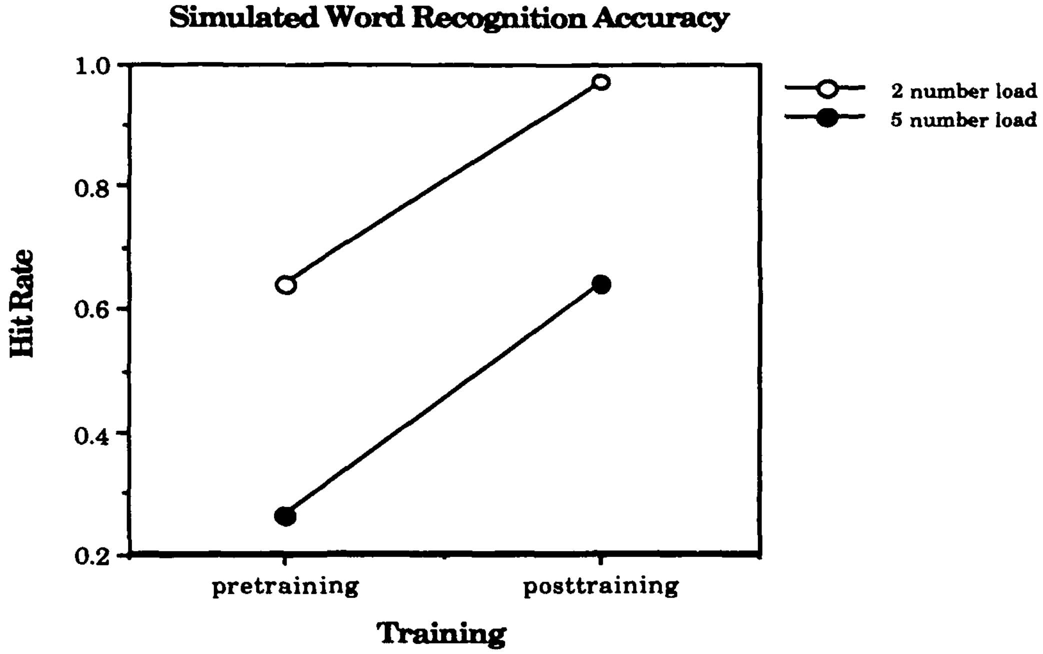

Figure 3.16. Simulated word recognition accuracy before and after training and at low and high levels of cognitive load.

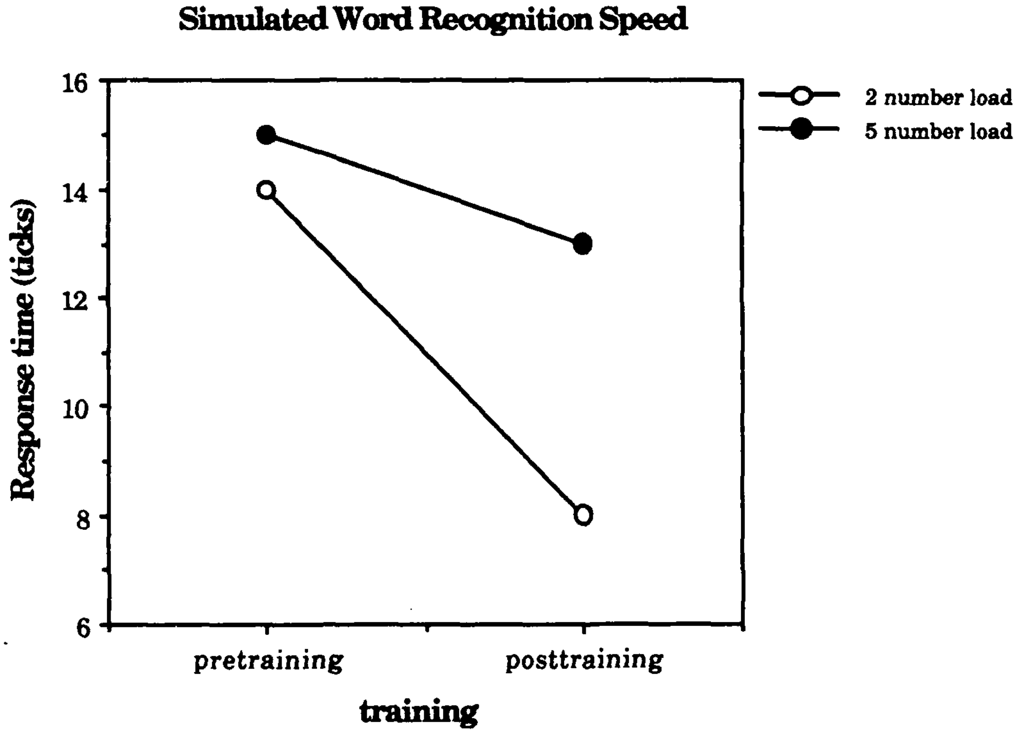

Figure 3.17. Simulated word recognition times before and after training, at high and low levels of cognitive load.

Furthermore, if we change only the data limitation by using more intelligible speech as we did in our third experiment, rather than affect the process limitation by training, the model produces additivity of response times just as we observed. Without training to optimize the allocation of capacity to the appropriate parts of the speech signal, the rate of information transfer from the signal stays the same.

This model represents a preliminary attempt to develop a formal account of the relation between learning and attention in speech perception that were originally described by Nusbaum and Schwab (1986). Although this model guided us in carrying out the third experiment, making the prediction that intelligibility alone would not interact with cognitive load, we still need to develop this model further. Specifically, we want to generalize this model to account for the relations between learning and attentional demands we see in perceptual normalization of talker differences. In addition, although this model provides a formal account of the results in our word monitoring studies with synthetic speech, we do not have a mechanism that can relate attention and learning and we intend to develop a connectionist model that will focus perceptual attention as a function of learning in accordance with the operating principles we have described here.

VI. Significance and Implications

Our program of research was directed at beginning investigations into the function and operation of attention in speech perception. The three areas addressed by this program of research concern: (1) the distribution of attention during recognition of speech and how attention is constrained by units of analysis and recognition; (2) how listeners accommodate the variability in the relations between acoustic and linguistic information that results from differences between talkers; and (3) how listeners learn to recognize synthetic speech by adaptively focusing attention on the acoustic properties of the speech. The findings of this program of research are very clear.

Listeners recognize speech as a continuous stream of phonetic information and do not recognize either syllables or syllable structure prior to phonetic recognition. Attention is distributed in time over the speech waveform processing earlier and later acoustic information without regard for syllable structure. Furthermore, our results, together with previous studies, suggest that speech is recognized using allophonic units (i.e., context-sensitive segments) rather than context-free phonemes or syllables. Our research also indicates that in mapping the speech of a particular talker onto these units, listeners employ two perceptual mechanisms. A structural estimation mechanism allows listeners to determine the linguistic representation that corresponds to a particular acoustic signal based solely on the information contained within the signal. However, it requires a substantial amount of attention and effort to compute this structural estimation. Contextual tuning allows listeners to learn the vocal characteristics of a talker based on this structural estimation, thereby reducing the attentional demands of speech perception. Finally, our research suggests the possibility that the mechanisms of structural estimation and contextual tuning may be more general than first thought. Instead of being processes that serve to normalize talker differences, these mechanisms may be fundamental to the processes of speech perception. Our research on the perception of synthetic speech demonstrates substantial learning effects occur in adult listeners who are long past the critical period of language acquisition. It is possible that when listeners first hear synthetic speech, they recognize it using structural estimation and therefore significant attention and effort is required. After moderate amounts of training, listeners may have used contextual tuning to focus attention on those aspects of acoustic-phonetic structure that are appropriate for the specific "synthetic talker." The focusing of attention may in fact modify the distribution of attention across phonetic context thus shifting the perceptual encoding of speech. The use of contextual tuning to modify the perceptual recognition of speech thus reduces the attentional demands of recognizing synthetic speech.

The results of these three projects represent an important beginning in our attempts to understand how attention operates in speech perception. It is interesting to note that theories of speech perception have always ignored or overlooked how attention and learning might be important in recognizing spoken language. In part this may be due to the influence of linguistics on the formation of these theories. Linguistics has defined a distinction between linguistic competence and linguistic performance. Competence refers to the knowledge of language a person might have, whereas performance refers to the use of that knowledge as limited by attention and memory. Theories of linguistics have treated performance limitations as irrelevant noise rather than a fundamental part of linguistic processing. Thus, the principles of cognitive psychology have been omitted from theories of speech perception (see Nusbaum, 1989). However, there is now an increasing body of evidence suggesting that it will be necessary to build theories of speech perception that are based on theoretical constructs of cognitive psychology including learning and attention (Nusbaum, 1989; Nusbaum & Schwab, 1986). In order to understand the mechanisms that mediate the perception and comprehension of spoken language, we will need to understand learning and memory, attention and effort, expectations, categories and similarity, and goals and plans are used in the processing of speech.

VII. References

Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in infancy. In M. M. Haith & J. J. Campos (Eds.), *Infancy and the biology of development*. Volume II of Carmichael's manual of child psychology, fourth edition. (P. H. Mussen, series editor). New York: Wiley and Sons, 1983.

Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford Science Publications.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation Vol. VIII*. New York: Academic Press.

Cutler, A., Butterfield, S., & Williams, J. N. (1987). The perceptual integrality of syllable onsets. *Journal of Memory and Language*, *26*, 406-418.

Egan, J. P. (1948). Articulation testing methods. *Larygoscope*, *58*, 955-991.

Fant, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.

Fudge, E. C. (1969). Syllables. *Journal of Linguistics*, *5*, 253-286.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: LEA.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, AV-16, 78-80.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibilityof eight text-to-speech systems. *Behavior research methods, instruments, & computers*, *18*, 100-107.

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (in press). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*.

Halle, M., & Vergnaud, J. R. (1980). Three-dimensional phonology. *Journal of Linguistic Research*, *1*, 83-105.

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

Luce, P. A., Feustel, T. C., & Pisoni, D. B. (1983). Capacity demands in short-term memory for synthetic and natural word lists. *Human Factors*, *25*, 17-32.

Martin, J. G., & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 473-488.

Moray, N. (1969). *Listening and attention.* Baltimore: Penguin.

Morin, T. M., & Nusbaum, H. C. (1988). Perceptual learning of vowels in a neuromorphic system. Submitted.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365-378.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology,7*, 44-64.

Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, *17*, 235-242.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Speech Perception Volume 1.* New York: Academic Press, 113-157.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab and H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Volume 1. Speech perception.* New York: Academic Press.

Nusbaum, H. C., Greenspan, S. L., & Pisoni, D. B. (1985). Perceptual attention in monitoring natural and synthetic speech. Paper presented at the 110th meeting of the Acoustical Society of America, Nashville, November, 1985. (also appears in *Research on Speech Perception Progress Report No. 12*, Speech Research Laboratory, Department of Psychology, Indiana University, 1986.)

Ohman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151-168.

Petersen, G. E., & Barney, H. L. (1952). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, *24*, 175-184.

Pisoni, D. B. (1978). Speech perception. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Volume 6. Linguistic functions in cognitive theory*. Hillsdale, NJ: Erlbaum.

Pisoni, D. B. (1981). In defense of segmental representations in speech processing. *Research on Speech Perception Progress Report No. 7.* Speech Research Laboratory, Indiana University.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 297-314.

Rand, T.C. (1971). Vocal tract size normalization in the perception of stop consonants. Haskins Lab. Status Rep. Speech Res. SR-25/26, 141-146.

Sawusch, J. R., Nusbaum, H. C., & Schwab, E. C. (1980). Contextual effects in vowel perception II: evidence for two processing mechanisms. *Perception & Psychophysics, 27,* 421-434.

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27,* 395-408.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086-1100.

Treiman, R., Salasoo, A., Slowiaczek, L. M., & Pisoni, D. B. (1982). Effects of syllable structure on adults' phoneme monitoring performance. *Research on Speech Perception Progress Report No. 8.* Indiana University, Speech Research Laboratory.

Treisman. A. M. (1969). Strategies and models of selective attention. *Psychological Review, 76,* 282-299.

Treisman, A. M., & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. *Cognitive Psychology, 14,* 107-141.

Verbrugge , R. R., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. *Language and Speech, 29,* 39-57.

Wood, C. C., & Day, R. S. (1975). Failure of selective attention to phonetic segments in consonant-vowel syllables. *Perception & Psychophysics, 17,* 346-350.