23980843

This is to certify that the

dissertation entitled

A Comparison of the Intelligibility

and Egronomics of Speech Synthesizers

presented by

Laura Jean Kelly

has been accepted towards fulfillment
of the requirements for

Doctor of Philosophy degree in Dept. of Audiology and Speech Sciences

Major professor

Date 28 October 1988

A COMPARISON OF THE INTELLIGIBILITY AND ERGONOMICS OF SPEECH SYNTHESIZERS

By

Laura Jean Kelly

A DISSERTATION

Submitted to
Michigan State University
in partial fulfullment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Audiology and Speech Sciences

1988

A Compariso

Five ex
speech synth
communicatio
accuracy by p
context gener
assessed sen
by the Speech
listening con
of passage c
performance
Experiment 4
communicatio
measured by a
sets of instr
Experiments 2

In Exper
presence of a
Experiment 2
result of spe
predictabili
between spee
for Experimen

ABSTRACT

A Comparison of the Intelligibility and Ergonomics of Speech Synthesizers

By

Laura Jean Kelly

Five experiments examined the intelligibility of five speech synthesizers and a human control at three points in a communication system. Experiment 1 assessed technical accuracy by performing spectral analysis of six vowels in CVC context generated by the speech sources. Experiments 2 and 3 assessed semantic precision using word recognition as measured by the Speech Intelligibility in Noise test (SPIN) and listening comprehension as measured by multiple choice tests of passage content. Two additional experiments assessed task performance via completion of oral instructions with (Experiment 4) and without (Experiment 5) options for communication repair. Completion of oral instructions was measured by a multiple instructions test (MIT) consisting of sets of instructions systematically varied in complexity. Experiments 2-5 employed listeners with normal-hearing.

In Experiments 2 and 3 stimuli were presented in the presence of a twelve-voice babble (+8 dB S/B) noise. Experiment 2 (N=12) revealed significant differences as a result of speech source for SPIN full-list, high- and low-predictability key word subtests and for the interaction between speech source and linguistic predictability. Results for Experiment 3 (N=12) revealed significant differences in

multiple c

Test compl

Stimu

presence o

Experiment

mean MIT s

levels of

complexity

significant

significant

The interac

was signifi

In Exp

from seven

the MIT. S

selected w

complexity

differences

5 revealed

A comp

Experiments

technical a

human archi

did not rea

multiple choice tests scores as a function of speech sources. Test completion time did not differ across sources.

Stimuli for Experiments 4 and 5 were presented in the presence of a twelve-voice babble noise (+10 dB S/B). In Experiment 4 (N=12), significant differences were seen between mean MIT scores as a function of speech source, complexity levels of the test and the interaction between source and complexity level. MIT item completion time did not differ significantly as a function of source, but did demonstrate significant differences as a function of complexity level. The interaction between speech source and item completion time was significant.

In Experiment 5 (N=14) subjects were allowed to select from seven communication repair options during presentation of the MIT. Significant differences among types of repair options selected were seen. The interaction between repair option and complexity level also was significant. A comparison of differences between MIT scores in Experiment 4 and Experiment 5 revealed scores in Experiment 5 to be significantly higher.

A comparison of the results of Experiment 1 to Experiments 2, 3 and 4 indicated the correlation between technical accuracy (as measured by summed vowel distances from human archival data) semantic precision and task performance did not reach the criterion for further analysis.

With love and greatest respect to my parents,

James and Joan Kelly,

who continue to be my teachers and friends.

I would like

Michael R. Chial

high standards.

like these with m

Special thanks to

equipment necessa

In addition

family members, E

contributed willi

critical points i

Along the w

fortunate to enc

who shared both

enriching the ex

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to Michael R. Chial for his unfailing patience and consistent high standards. Beyond technical and professional skills, I take these with me as examples of the best in teaching. Special thanks to Lonnie Smrkovski for arranging access to equipment necessary for completion of this work.

In addition, I extend my love and gratitude to my family members, Bill, Sandy, John, James and Patrick who contributed willing hands and immeasurable support at several critical points in the research process.

Along the way to reaching this goal, I also was fortunate to encounter a special group of doctoral students who shared both crisis and celebration. Thank you for enriching the experience with your help and friendship.

CHAPTER

TABLE OF CONTENTS

CHAPTER Page

METHODS (continued)

Table

LIST OF TABLES

Table

Table Page

Table

Table Page

Table | Page

Table

Table Page

Table Page

Table	Page

Figure

LIST OF FIGURES

Figure (Co

3.2. Mean
for S
word
repre
teste
stand

3.3. Mean
standa
as a
obser
monau
standa

3.4. Mean a
standa
and lo
speech
12 nor
lower

3.5. Illust
compar
percen
and lo
speech
connec

3.6. Mean p
for mu
of spe
12 nor
lower

3.7. Mean a
standa
functi
observ
monaur
standa

3.8. Mean m
standa
speech
of 12
lower

3.9. Illust
compar
compre
connec

Figure (Cont.) Page

Figure

Figure (Cont.) Page

Figure

Figure

Figure (Cont.) Page

Figure (Cont.)

D-3. Mean perce
SPIN high-
function o
speech sou
normal-hea
histogram

D-4. Mean arcsi
standard d
predictabi
signal-to-
source. E
normal-hea
lower hist

D-5. Mean perce
SPIN high-
function o
speech sou
normal-hea
histogram

D-6. Mean arcsi
standard d
predictabi
signal-to-
source. E
normal-hea
lower hist

Figure (Cont.) Page

Digital spe

devices used i

augmentative c

proliferation

effects of mac

Assuming the d

make more effe

is essential t

upon the perfo

appropriate co

can be underta

Of particu

of a task-appr

intelligibilit

speakers and t

acoustic envir

dramatically i

required and t

presence of co

interact to in

complexity, ta

message, limit

listener exper

Chapter I

BACKGROUND

Introduction

Digital speech synthesizers have become commonplace in devices used in business, industry, and education, and as augmentative communication aids to the handicapped. This proliferation raises difficult questions regarding the effects of machine-generated speech upon communication. Assuming the decision to use synthetic speech is intended to make more effective use of human communication resources, it is essential to know the nature and degree of these effects upon the performance of tasks. Once these are determined, appropriate cost/benefit evaluations of available systems can be undertaken.

Of particular interest in the development and selection of a task-appropriate speech synthesis device is the intelligibility of the system in comparison to both human speakers and to other synthetic speech systems. The various acoustic environments in which these systems are used can dramatically influence the degree of intelligibility required and the robustness of intelligibility in the presence of competing signals. A wide variety of variables interact to influence intelligibility including signal complexity, task complexity, linguistic structure of the message, limitations of the human processing system and listener experience (Nusbaum and Pisoni, 1985). Measurement of intelligibil

accuracy, prec

An effecti

associated wit

communication

and Weaver (19

triad of issue

(1) The ac
commun

(2) The pr
convey

(3) The ef
affect

Figure 1.

based on the

1981), and C

the transfer

thoughts) fro

the process b

wishes to tra

The purpose o

this example,

source to eff

defined as an

signal, thus

response. One

information t

signal source

of intelligibility requires attention to the issues of accuracy, precision, sufficiency, and utility.

An effective means of organizing the many concerns associated with the evaluation of synthesized speech is a communication model such as that model developed by Shannon and Weaver (1949). The foundation of their approach is a triad of issues described as follows:

(1) The accuracy with which the symbols of communication are transmitted (technical accuracy).

(2) The precision with which the transmitted symbols convey the desired meaning (semantic precision).

(3) The effectiveness with which the received meaning affects conduct in a desired way (task performance).

Figure 1.1 offers an idealized model of speech synthesis based on the work of Shannon and Weaver (1949), Flanagan (1981), and Chial (1986). Communication can be described as the transfer of information (i.e., facts, feelings, thoughts) from one place to another. In the present model, the process begins with an information source (human) who wishes to transfer information to a receiver (also human). The purpose of communication is variable; but is assumed, in this example, to be an intent on the part of the information source to effect a response from the receiver. Noise can be defined as anything which increases the ambiguity of the signal, thus reducing the likelihood of the desired receiver response. One means for generating an auditory signal for information transfer is a speech synthesizer. For this signal source to function, however, information must be


Components of a Speech Synthesizer
Information Source
Encoding
Text
Text to Symbol Conversion
Symbol to Output code Processor

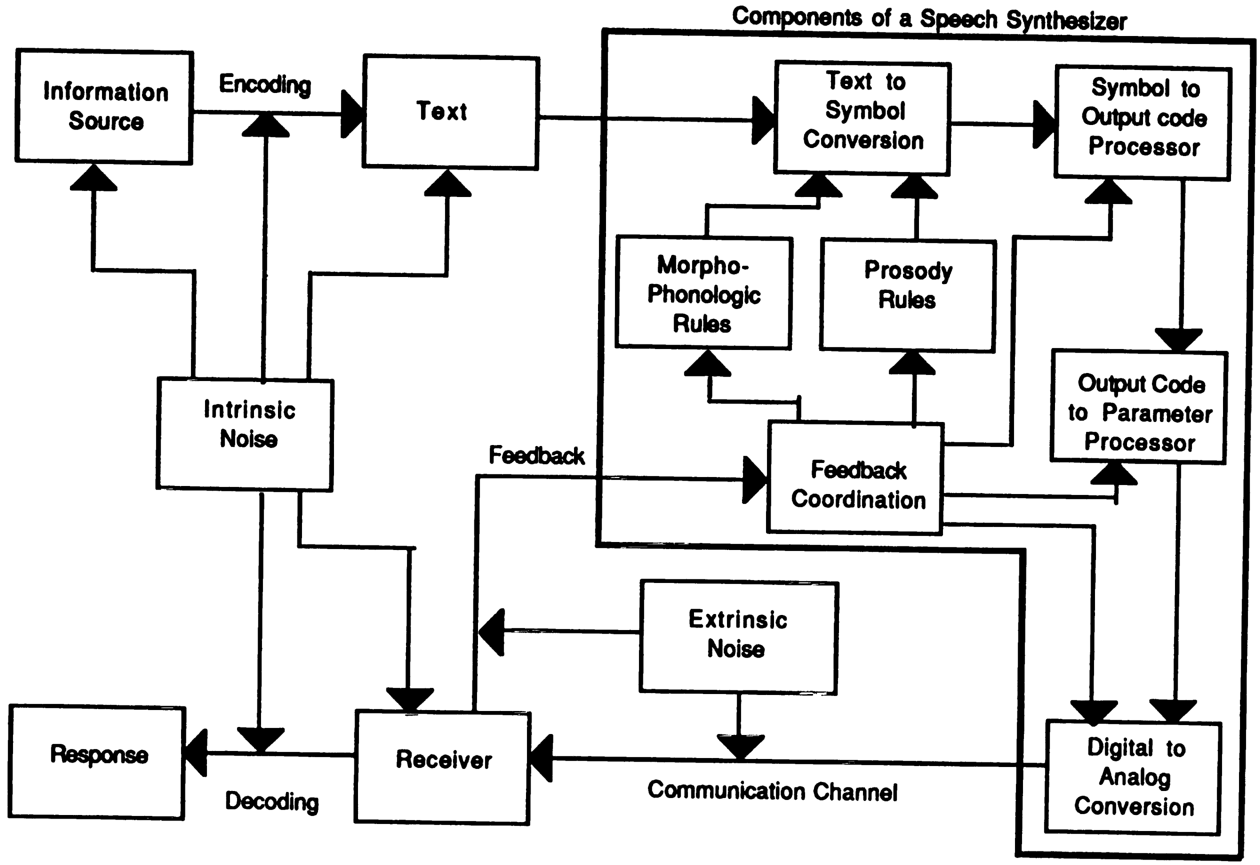


Figure 1.1. Idealized model of an interactive communication system containing a speech synthesizer.

translated or e

encoding proces

the communicati

e.g encoding a

acoustic signal

include vocabul

generation. Ex

or traffic noise

Text, includ

suprasegmental i

phonetic informa

synthesizer. S

further recoded

speech wave. A

output the synt

through a commu

A feedback

control of the

accuracy of fee

of such interac

efficiency and

directed commun

effectiveness o

in the amount o

role and use o

as an ergonomi

ergonomics.

translated or encoded into a form the device can use. This encoding process is the first potential source of noise in the communication system. Noise sources can be intrinsic (e.g encoding and decoding) or extrinsic e.g.(a competing acoustic signal). Examples of intrinsic noise sources include vocabulary selection and the accuracy of text generation. Extrinsic noise might include competing talkers or traffic noise.

Text, including punctuation marks that cue suprasegmental information, is recoded into segmental phonetic information through of algorithms stored in the synthesizer. Symbols generated from these algorithms are further recoded into parameters used to create a digitized speech wave. A digital-to-analog converter is used to output the synthesized wave for transmission along along or through a communication channel.

A feedback component may exist to permit interactive control of the synthesizer. The quality, frequency and accuracy of feedback from the receiver and the utilization of such interaction by the source can influence the efficiency and effectiveness of communication. In goal-directed communication, variations in the efficiency and effectiveness of the communication process cause variations in the amount of work required to accomplish the goal. The role and use of human energy in such tasks can be addressed as an ergonomic issue, specifically communication ergonomics.

Interaction

into three mode

This rubric (hi

technology) is

communication b

more humans emp

communication.

As with the

assumed in which

channel of some

machine. "Chann

transmissive med

media (print, vi

The simplex

of information i

always the sourc

(see Figure 1.2)

involving human

of music. The h

"turns" in a dis

role of sender a

debates, serious

answering system

actors simultane

serving as both s

communication in

arguments. (Chia

Interactions between humans and computers can be divided into three modes: simplex, half-duplex and full-duplex. This rubric (historically applied to serial communication technology) is useful for describing the nature of communication between human and computer, or between two or more humans employing a computer as a mediator of communication.

As with the previous communication model, a system is assumed in which two actors exchange information over a channel of some sort. The actors may be either human or machine. "Channel" in this context refers to any transmissive medium (acoustical, optical, electronic) or media (print, video, film).

The simplex mode of interaction involves transmission of information in only one direction, that is, one actor is always the source, the other actor is always the receiver (see Figure 1.2). An example of simplex communication involving human actors is a taped or phonographic recording of music. The half-duplex mode allows actors to take "turns" in a discrete non-overlapping manner, exchanging the role of sender and receiver. Examples include formal debates, serious telephone conversations and telephone answering systems. The full-duplex mode is one in which actors simultaneously engage in bidirectional communication, serving as both sender and receiver. Examples of full-duplex communication include lively conversation and impassioned arguments. (Chial, 1984)

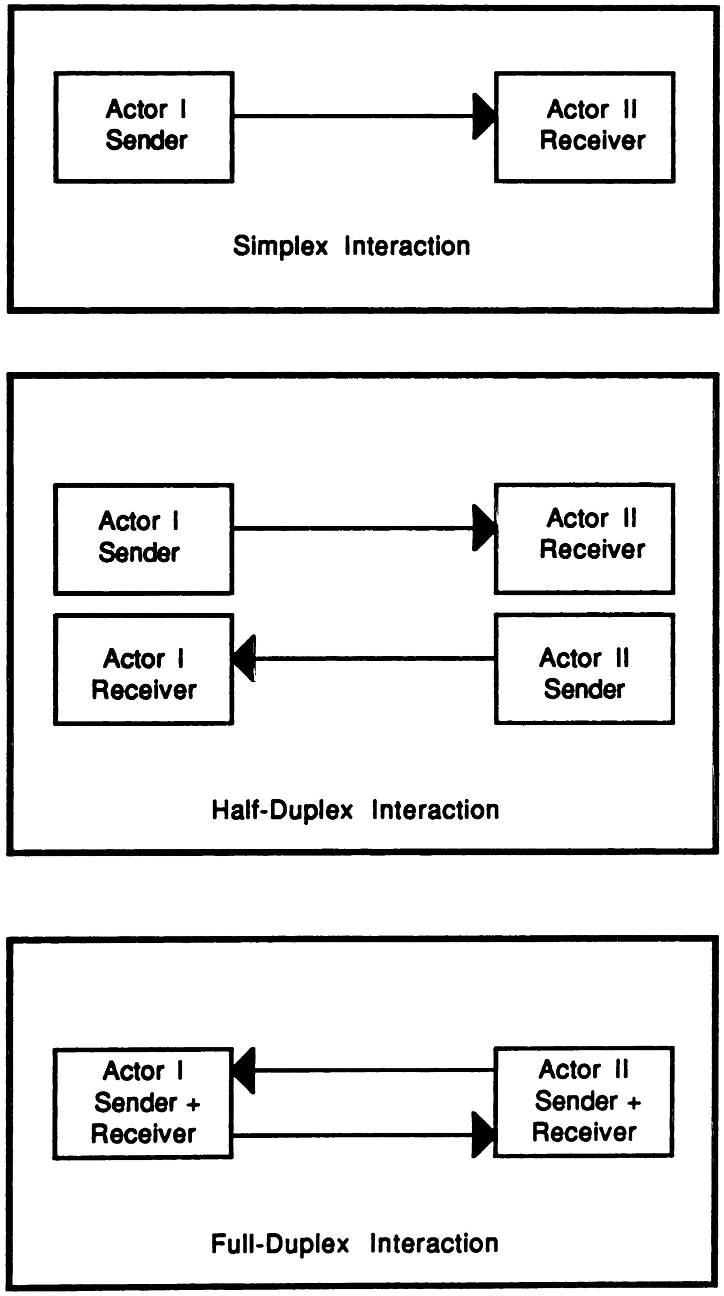


Figure 1.2 The three categories of human-computer interaction.

It is possi
any point in th
communication.
the selection c
of the device c
process most cr
To paraphrase S
practical to de
everything perf
handle well the
and should rede
task (1949, p.

Table 1.1 s
synthesized spe
receptive inte
or sentence re
This approach
accuracy of th
attempted to c
quality and ex
intelligibilit
attempts have
focusing on te
performance.

The goal c
a group of spe
system levels c

It is possible to study aspects of synthesized speech at any point in the communication system or through any mode of communication. A practical approach to evaluation lies in the selection of a method appropriate to the task required of the device or for that portion of the communication process most critical to performance in a given situation. To paraphrase Shannon and Weaver, if it is not possible or practical to design an evaluation approach which can handle everything perfectly, then a system should be designed to handle well the jobs it is most likely to be asked to do, and should redesign itself to be less efficient for the rare task (1949, p. 14).

Table 1.1 summarizes the approaches used to assess synthesized speech. Most prior research has concentrated on receptive intelligibility, usually through the use of word or sentence recognition and listening comprehension tests. This approach provides information limited to the semantic accuracy of the speech synthesis systems. A few studies also attempted to compare synthesizers on the basis of perceived quality and explored the relationship between intelligibility and perceived naturalness. To date, no attempts have been made to design and evaluate procedures focusing on technical precision or on complex task performance.

The goal of this study was three fold: (1) to evaluate a group of speech synthesizers at the three communication system levels of technical accuracy, semantic precision and

TYPES OF STIMULI

Phonemes	Words			Sentences						Connected Discourse	
						[illegible]	HAP	AID	SV	ARC	MISC.

TYPES OF STIMULI

Speech Synthesis Systems Evaluated	Phonemes: CV	Words: MRT	Words: W-22	Sentences: ESWS	Sentences: PSS	Sentences: HAS	Sentences: HAR	Sentences: AID	Sentences: SV	Connected Discourse: ARC	Connected Discourse: MISC.
MITalk-79		R(1) S(5) K(2)				R(1)	R(1)			R(3)	
TSI Prototype-1 (of Prose 2000)		A(1,4)				A(1)	A(1)	N(1,9)		A(3)	
DECtalk v 1.8 (Paul)	O(2)	D(1,4) J(2)			J(1) K(2)			N(1,9)	L(3)		
DECtalk v 1.8 (Betty)		C(1,4)							L(3)		
Infovox SA 101		C(1,4)			J(1) K(2)	K(2)	K(2)		L(3)		
Prose 2000 v 3.0		C(1,4) K(2)			J(1) K(2)				L(3)		
Berkely		C(1,4)									
Echo II		C(1,4)		G(1)				I(1,6)			
Echo II+								N(1,9)			
Votrax Type N' Talk	O(2) E(8)	C(1,4)	E(8) T(8)	G(1)		T(8) E(8)	T(8) E(8)		L(3)		T(8) F(2) H(1)
Votrax Personal Speech System								I(1,6) N(1,9)			
Misc.			B(1,5,7,8)						Q(3)		P(2) F(2) M(2)

Table 1.1 Summary of speech synthesis systems evaluated, stimuli employed and assessment goals of the research reviewed.

CV
MRT
W-22
ESW
PSS
HAS

LEGEND FOR TABLE 1.1

STIMULUS MATERIALS

CV	Consonant Vowel Combinations	HAR	Harvard Psychoacoustic Sentences
MRT	Modified Rhyme Test	AID	Assessment of Intelligibility of Dysarthric Speech
W-22	CID W-22	SV	Sentence Verification
ESWS	Experimenter Selected Words + Sentences	ARC	Adult Reading Comprehension Tests
PSS	Phoneme Specific Sentences	MISC.	Selected Passages
HAS	Haskins Laboratories Anomalous Sentences		

ASSESSMENT GOALS

1	Word Recognition	5	Effect of Noise
2	Speech Quality	6	Effect of Rate
3	Listening Comprehension	7	Effect of SPL
4	Open vs Closed Set	8	Effect of Training
		9	Effect of Age

AUTHORS

A	Bernstein & Pisoni (80)	K	Logan & Pisoni (86b)
B	Chial (76)	L	Manous, Pisoni, Dedina & Nusbaum (85)
C	Greene, Logan & Pisoni (86)	M	McHugh (76)
D	Greene, Manous & Pisoni (84)	N	Miranda & Beukelman (87)
E	Greenspan, Nusbaum & Pisoni (86)	O	Nusbaum, Greenspan & Pisoni (86)
F	Hersch, & Tartaglia (81)	P	Nye, Ingemann, & Donald (75)
G	Hoover, Reichle, Van Tasell, Cole (87)	Q	Pisoni & Dedina (86)
H	Jenkins & Franklin (81)	R	Pisoni & Hunnicutt (80)
I	Kraat & Levinson (84)	S	Pisoni & Koen (86)
J	Logan & Pisoni (86a)	T	Schwab, Nusbaum & Pisoni (85)

task perfo
techniques
process to
occur as a
initial da
employing

Klatt
synthesis
converting
an ASCII (
interchang
into the s
binary val
marks and
Klatt (198
the follow

(1) Re
cha

(2) Se
str

(3) As
str

(4) Det

(5) Ass

At the pres
perform a s
On this bas

task performance; (2) to apply a combination of behavioral techniques used at the same level of the communication process to determine whether different rankings of systems occur as a result of different approaches; and (3) to obtain initial data on the role of communication repair in systems employing synthetic speech.

Text-to-Speech Conversion

Klatt (1987) discussed the state of the art in speech synthesis technology. The first step in the process of converting text to an auditory signal is the assignment of an ASCII (American standard code for information interchange) code to each typed character or string entered into the synthesizer. ASCII codes are 7-bit or 8-bit binary values assigned to letters, numbers, punctuation marks and special characters (Chial, 1984). According to Klatt (1987) the resulting code is then ideally subjected to the following analysis:

(1) Reformat all digits, abbreviations and special characters into words and punctuation.

(2) Section sentences to establish a surface syntactic structure.

(3) Assign a stress pattern appropriate to the surface structure.

(4) Determine a phonemic representation for each word.

(5) Assign a stress pattern to each word.

At the present time, text-to-speech systems are unable to perform a semantic analysis and thus assign stress patterns on this basis. Instead systems with the option of sentence

level into
pattern wh
such as co
patterns a
user codes
Therefore,
representa
This is ac
entries in
listed are
suffixes a
to find a
pronunciat
dictionari
deal with
Figure 1.3
conversion

Rule u
accordance
considered
These syst
underlying
comparisons
Yet enough
between rul
empirical c

level intonation patterns use a "generic" inflectional pattern which may or may not change with punctuation markers such as commas and question marks. No additional stress patterns are inserted on the basis of semantics unless the user codes the input with additional stress markers. Therefore, the system proceeds to the derivation of phonemic representation and stress assignment at the word level. This is accomplished via a word-by-word comparison to entries in a pronunciation dictionary. Those words not listed are broken into pieces to remove prefixes and suffixes and compared again. If the system is still unable to find a match for the root word rules for letter pronunciation are used. Some systems incorporate dictionaries to check for exceptions to stress rules or to deal with special cases such as proper names (Klatt, 1987). Figure 1.3 reproduces Klatt's (1987) model of text-to-speech conversion.

Rule used for text-to-speech translation vary in accordance with speech synthesis systems and can be considered the primary determiner of perceptual differences. These systems are often proprietary, making analysis of underlying rule structures difficult. Even if detailed comparisons of rule structures were possible there is not yet enough information on the cause-effect relationship between rules configuration and perception. Consequently, empirical comparisons of rule structure are necessary.

Figure 1.
orthograp
(Klatt, 19

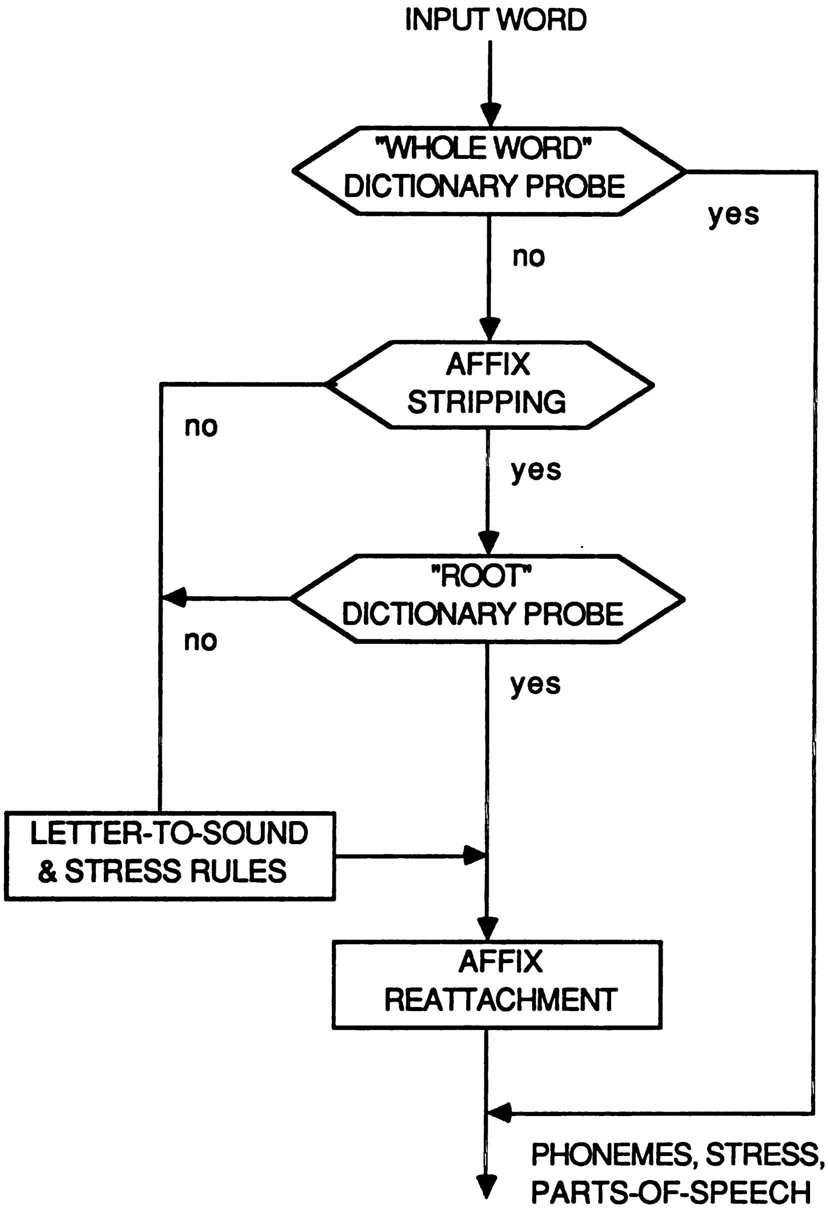


Figure 1.3. The steps involved in converting as ASCII orthographic representation for a word into phonemes. (Klatt, 1987, p.768)

Miller (198

of speech can b

perceptual spac

normally occurr

compared to tar

within the stor

particular elem

be accomplished

characteristics

measurements of

frequencies. S

characteristics

plotting the ph

used a three di

similarities am

dimensions. Mi

elements should

present in the

shapes as chara

relationships.

these dimension

characterized b

position along

allows for simp

dimension witho

by the signal.

Speech Perception

Miller (1984) proposed that short-term spectral patterns of speech can be represented as points in a "auditory-perceptual space". Speech spectra are integrated with normally occurring silences to form a pattern which is compared to targets learned over time. If the pattern falls within the stored target zones it will be perceived as a particular element. Description of a phonetic pattern can be accomplished using the spectral characteristics. Vowel characteristics traditionally have been quantified through measurements of fundamental frequency and formant frequencies. Some researchers have used these characteristics to describe differences in categories by plotting the phonemes along two dimensions. Shepard (1972) used a three dimensional plot to demonstrate that perceptual similarities among vowels tend to line up along formant dimensions. Miller theorized perceptual coding of phonetic elements should not be based on the number of prominences present in the spectrogram, but the pattern of the spectral shapes as characterized by log-power and log-frequency relationships. He suggested that vowels are plotted using these dimensions, the sensory perception of the phoneme is characterized by "spectral shape" as opposed to absolute position along any one dimension. This reliance on shape allows for simple transposition of the vowel along either dimension without altering the phonetic information carried by the signal.

Accordin

as a single

the characte

z= log (F_2/F

defined as th

by a constant

F_3 are freque

prominences.

space, allowi

based upon di

semitones (Mi

Although

design of coc

prostheses wi

information t

the character

here that his

as a means of

between human

basis of the c

three dimensio

By far th

synthesized sp

intelligibilit

encompassing th

According to Miller, "spectral shape" can be represented as a single point plotted in three-dimensional space with the characteristics of x= log (F_3/F_2), y= log (F_1/F_0') and z= log (F_2/F_1). In the case of periodic speech, F_0' is defined as the fundamental frequency of the voice multiplied by a constant (1.5 times greater for males), and F_1, F_2 and F_3 are frequency locations of the first three spectral prominences. Vowels can be plotted in three-dimensional space, allowing for the calculations of class differences based upon distance as measured in octaves, cents or semitones (Millers, 1982)

Although Miller (1984) directed his interest to the design of cochlear implants, he speculated that cochlear prostheses will only succeed in transmitting phonetic information to the degree they are successful in matching the characteristics of spectral envelopes. It is suggested here that his model also can be applied to the signal source as a means of indexing technical accuracy. Differences between human and synthetic speech can be described on the basis of the degree of separation among points plotted in three dimensional space.

Speech Understanding

Introduction

By far the most popular technique for assessing synthesized speech has been speech understanding or intelligibility. Intelligibility is defined here as encompassing the discrimination, recognition and

comprehens
defined as
speech sig
writing. W
isolation o
reproductio
on the othe
speech sig
different f
tests of se
recognition
tasks using
in their ef

Word Recogn

Chial (
recognition
and normal
Auditory Te
earphones.
in quiet at
addition, th
was investig
was obtained
score of alm
improvement
for the spee

comprehension of speech stimuli. Speech recognition is defined as the process by which an individual receives a speech signal, then immediately reproduces it verbally or in writing. Word-recognition tasks may involve word in isolation or in sentences. Scoring is based on the correct reproduction of single words. Speech comprehension tests, on the other hand, generally require longer retention of the speech signal as well as recognition of message content in a different form (usually written). Researchers have employed tests of segmental intelligibility, word-recognition, word-recognition in sentences, as well as listening comprehension tasks using sentence verification and continuous discourse in their efforts to compare systems.

Review of the Literature

Word Recognition

Chial (1976) designed four experiments to evaluate word recognition using the VOTRAX VI phonetic speech synthesizer and normal hearing listeners. The stimuli consisted of CID Auditory Test W-22, List 1 presented monaurally under earphones. Experiment 1 was designed to assess performance in quiet at comfortable listening levels (70 dB SPL). In addition, the effect of repeated trials on word recognition was investigated. An average word recognition score of 55% was obtained for the VOTRAX VI, in comparison to an average score of almost 100% for the human control. A significant improvement in performance was noted with repeated trials for the speech synthesizer.

Experimen
sensation lev
reception thr
differences i
levels, with
definite plat
results were
talker.

Experiment
six different
dB for the VO
SPL of 70 dB
recognition s
with significa
plateau was re
considered cor
human talker.

Experiment
when human and
lists were pre
no improvement
significant di
results of exp
accelerated le
presentation o

To investig
synthesized spe

Experiment 2 measured word recognition at six different sensation levels (SL) from +5 dB to +30 dB (re: Speech reception threshold) for the VOTRAX VI only. Significant differences in performance were noted among sensation levels, with a possible plateau noted at +20 dB and a definite plateau seen at +30 dB. Chial (1976) noted these results were similar to what would be found with a human talker.

Experiment 3 assessed word recognition performance at six different signal-to-noise ratios (S/N) from -5 dB to +20 dB for the VOTRAX VI only. The stimuli were presented at a SPL of 70 dB in the presence of white noise. Word recognition scores improved as the S/N became more positive, with significant differences noted at all S/N until a plateau was reached at +15 dB. These findings were considered consistent with what would be expected with a human talker.

Experiment 4 investigated the effect of repeated trails when human and synthesized speech are alternated. Ten test lists were presented at SPL of 65 dB. The results revealed no improvement in performance for the human talker, but significant differences for synthesized speech. Comparing results of experiments 1 and 4, Chial reported an accelerated learning effect associated with alternating presentation of talkers.

To investigate the effects of noise on the perception of synthesized speech, Pisoni and Koen (1981) presented

materia
three d
The Mod
SPL of
for eac
respons
results

Pisc
synthesi
speech.
may inte
effects
Cole (19
word-rec
the VOTR
consonan
place, m
in eithe
(p.30).
were gen
set was
"Say the
second s
provided
correctl

Resu
synthesi

material generated by the MITalk text-to-speech system at three different speech- to-noise ratios (+30,+20, and 0 dB). The Modified Rhyme Test (MRT) was presented at an average SPL of 80 dB with white noise attenuated to meet criteria for each noise condition. In addition, both open and closed response modes were employed. Table 1.2 summarizes the results.

Pisoni and Koen concluded that the intelligibility of synthesized speech is affected more by noise than is human speech. It also was suggested that this signal distortion may interact with different processing tasks to produce effects on intelligibility. Hoover, Reichle, Van Tasell and Cole (1987) compared single word-recognition scores and word-recognition in sentences generated by the Echo II and the VOTRAX Type 'N Talk to human speech. Twenty seven consonant-vowel-consonant words were selected such that "all place, manner, and voicing characteristics were represented in either the initial or final positions of the word" (p.30). For the contextual task, two sets of sentences were generated with these words in the final position. One set was designated as low-probability and used of the phrase "Say the word " as the precursor to the stimulus item. The second set, designated as high-probability sentences, provided sufficient context for 90% of 15 listeners to correctly guess the item from the sentence content.

Results revealed significant differences between each synthesizer and human speech on the basis of the percentage

Table
ratios

Source
MITalk
Human

Table 1.2. MRT scores (%-correct) at three speech-to-noise ratios as a function of speech source.

	Closed Set			Open Set		
Source / Sp/N	+30	+20	0	+30	+20	0
MITalk	93	89.4	56.6	79	73.5	28.9
Human	99	97.2	69.5	92	88.9	40.3

(Adapted from Pisoni and Koen, 1981)

of wor

senten

signif.

catego

when th

analys

recogn

such p

Analys

substi

initial

1% of t

items w

subject

or aff

VOTRAX.

Vis

with th

between

positi

wave s

the co

resemb

corres

subjec

of words correctly identified in isolation and within sentences. Between synthesizers the VOTRAX performed significantly better than the Echo on both sentence categories. No difference was seen between the synthesizers when the words were presented in isolation, but, an error analysis revealed a substantial difference in the recognition of stop consonants. The Echo II proved to be much poorer (23% correct) than the VOTRAX (52% correct). Analysis of subject response patterns indicated 75% substituted the phoneme /m/ for stop consonants in the initial position (usually replacing /b/), whereas less than 1% of the subjects responded with this phoneme for the same items when presented by VOTRAX. In the final position, subjects identified 20% of the Echo II items as fricatives or affricates compared to 1% of the items presented by VOTRAX.

Visual inspection of the acoustic waveforms associated with these stop consonants revealed distinct differences between the two synthesizers. The /b/ in the initial position produced by the Echo II begins with a low-frequency wave similar to a nasal consonant. In the final position the consonant release is followed by an aspiration resembling a fricative. Thus, the acoustic features correspond with many of the perceptual errors made by the subjects.

Word Recognition In Sentences

Pisoni and Hunnicut (1980) investigated the intelligibility of the MITalk using a three-phase approach consisting of segmental intelligibility test, a word recognition test and a listening comprehension test. Segmental intelligibility was evaluated with the Modified Rhyme Test (MRT) under earphones. Normal hearing subjects (N=72) produced an overall error rate of 6.9% with scores of 4.6% and 9.3% for consonants in the initial and final positions, respectively. Nasals were found to have the highest error rate (27.6%). An overall error rate of .6% was obtained when human speech was used.

Word recognition in sentences was evaluated using the Harvard Psychoacoustics Laboratory Sentences (Egan, 1948) and semantically anomalous sentences created at Haskins laboratory (Nye and Gartenby, 1974). Scores of 93.2% and 78.7% correct were obtained for synthetic speech, while scores of 99.2% and 97.7 % were seen for recordings using natural speech.

Listening comprehension was assessed using narrative passages selected from adult reading comprehension tests. Three groups of subjects were employed. One group listened to MITalk and a second to human speech. A third group viewed the passages in typed form. A set of multiple-choice questions was administered immediately following the task. Scores of 70.3%, 67.2% and 77.2% were obtained for the three groups, respectively.

In
System
a huma
batter
differ
measur
were s
the an
speech

Gr
invest
synthe
provid
this s
"Beau
MRT as
senten
stimul
earpho
broadb
error
voice)
contra
13.2%
respec
demons
9.5% a

In a study comparing the MITalk to the Telesensory Systems text-to-speech device (a system based on MITalk) and a human control, Bernstein and Pisoni (1980) used the some battery of tests just described. No significant differences were reported among the systems on any of the measures employed. The largest differences and error rates were said to occur on the sentence materials, particularly the anomalous sentences. Specific percentages for each speech and test were not reported.

Greene, Manous and Pisoni (1984) conducted an investigation using the Digital Equipment Corporation speech synthesis system version 1.7 (DECtalk). This device provides six different voices, two of which were used for this study: one male ("Perfect Paul") and one female ("Beautiful Betty"). Once again, the investigators used the MRT as a test of segmental intelligibility the Harvard PAL sentences and the Haskin anomalous sentences materials as stimuli. The materials were presented to the subjects under earphones (80 dB SPL) in the presence of 55 dB SPL of broadband noise to mask tape hiss. The results showed an error rate on the MRT of 3.3% (male voice) and 5.6% (female voice) when a closed-set response format was used. In contrast, the open-set format resulted in an error rate of 13.2% and 17.5% for the male and female voices, respectively. The Harvard and Haskins sentence materials demonstrated error rates of 4.7% and 13.2% (male voice) and 9.5% and 24.0% (female voice).

A co

in a sen
(1986).
Specific
sentence
of words
the sent
voiced f
sentence
on the b
An error
incorrec
basis of
A signif
category
between
differen
findings
as the m
that the
as evide
all sour
phonetic
about sy
of differ

Kraat
for sente

A comparison of the intelligibility of phoneme classes in a sentence context was undertaken by Logan and Pisoni (1986). The stimuli consisted of a subset of the Phoneme Specific Sentences (Huggins and Nickerson, 1985). These sentences are designed such that each item contains a number of words from a specific class of phonemes. For example, the sentence "Those waves veer over", contains numerous voiced fricatives. The subjects were asked to transcribe sentences presented under earphones. The items were scored on the basis of omissions, transpositions, and additions. An error in any category meant the sentence was counted as incorrect. Analysis revealed significant differences on the basis of synthesizer and on the basis of phonemic category. A significant interaction between voice and phonetic category also was noted. Significant differences were seen between DECtalk and Prose versus Infovox; however, no difference was seen between DECtalk and Prose. These findings are in contrast to previous research using the MRT as the measure of intelligibility. The authors concluded that the Phoneme Specific Sentence is a more difficult test as evidenced by the higher overall error rates exhibited by all sources. They suggested that errors at the level of phonetic categories can reveal more precise information about synthesizers than sources of error even in the absence of differences in overall performance among synthesizers.

Kraat and Levinson (1984) compared the intelligibility for sentences of the Echo II and the VOTRAX Personal Speech

Sys

1 2

con

Int

Bei

Lev

the

int

Six

con

wri

lou

in

PSS

and

pro

of

usi

pro

Jud

del

Sti

(19

thes

the

Beu

System (PSS) produced at (1) normal rates and (2) with a 2 1/2 second pause between each word. Test materials consisted of 64 sentences from the Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). Although not so stated by Kraat and Levenson, these materials originally were designed to assess the extent of motor speech difficulties upon the intelligibility of speech (Yorkston and Beukelman,1981) Sixteen sentences were randomly assigned to each of four conditions. Twenty normal-hearing adults were asked to write the sentences following their presentation by loudspeaker. The results revealed a significant difference in performance with percent-correct scores of 70.4% for the PSS and 45.7% for the Echo II in normal conditions and 84.3% and 81.1% in the pause conditions. The pause condition provided significant improvement for both synthesis devices.

Kraat and Levinson (1984) also evaluated the adequacy of the text-to-speech conversion rules of the two systems by using five speech pathology graduate students as judges of pronunciation of 1500 words produced by the synthesizers. Judges determined whether syllables had been added or deleted and whether vowel substitutions had occurred. Stimulus items were taken from the Thorndike and Lorge (1944) list of the 1000 most frequent English words. Of these, 45 were judged as mispronounced by the PSS and 175 by the Echo II. An additional 500 words were taken from the Beukelman, Yorkston, Poblete, and Naranjo's (1984) lexicon

of w
corr
misp
to h

of s
syst
(see
rema
these
discu
repor
forma
forma
T
on th
high-
MITal
SA101
synthe
catego
speech
Mo
five
System
"Kit th
single

of words commonly employed by users of augmentative communication devices. The PSS was judged to have mispronounced 36 of these words while the Echo II was judged to have mispronounced 55 of the items.

Greene, Logan and Pisoni (1986) summarized error rates of segmental intelligibility for eight speech synthesis systems from experiments conducted over the past seven years (see Table 1.3) The subject criteria and procedures remained the same for all the systems evaluated. Two of these systems DECtalk and MITalk, already have been discussed here. As with the previous studies, data were reported for an open-set response format (open-set response formats produced higher error rates than closed-set response formats).

These authors suggested a four level grouping of devices on the basis of intelligibility: (1) natural speech, (2) high-quality synthetic speech (DECtalk, Prose 3.0 and MITalk), (3) Moderate-quality synthetic speech (Inovox SA101, Berkely, and TSI proto-1), and (4) low-quality synthetic speech (VOTRAX Type 'N Talk and Echo). These categories reflect the effectiveness of rules for text-to-speech conversion used in each system.

More recently, Mirenda and Beukelman (1987) compared five synthesized voices (Echo II+, VOTRAX Personal Speech System, and DECtalk; "Perfect Paul", Beautiful Betty" and "Kit the Kid") and a human speaker (female) using both single word and word recognition in sentences. Speech

Tabl
cons

Voice
Natur
DECta
DECta
MITal
Pross
Infov
Berke
TSI-p
VOTRA
Echo

Table 1.3. MRT Error Rates (%) overall and error rates for consonants in initial and final positions.

Voice	Error Rate (%) Inital	Final	Overall
Natural Speech	0.59	0.56	0.53
DECtalk 1.8 Paul	1.56	4.94	3.25
DECtalk 1.8 Betty	3.39	7.89	5.72
MITalk -79	4.61	9.39	7.00
Pross 2000 V3.0	7.11	4.33	5.72
Infovox SA 101	10.00	15.00	12.50
Berkely	9.78	18.50	14.14
TSI-Protosype I	10.78	24.72	17.75
VOTRAX Type'n'Talk	32.56	22.33	27.56
Echo	35.56	35.56	35.56

(Greene, Logan, and Pisoni, p. 104, 1986)

stimu
synth
phone
and s
Compu
Speech
Twelve
source
age gr
ages
were p
Record
tape e
levels
report
trial
presen
experi

differ
differ
indica
word re
signif
as well
indica
designe

stimuli also were generated for the Echo II+ and VOTRAX synthesizers using both standard English spelling and phonetic coding. Stimuli were a pool of single words (600) and sentences (1,100) selected randomly using the Computerized Assessment of Intelligibility of Dysarthric Speech (CAIDS: Yorkston, Beukelman, & Traynor,1984). Twelve sentences and fifty words were used with each speech source. The subjects consisted of five listeners from three age groups; adult (ages 26 -40), older elementary children (ages 10-12) and younger elementary children (6-8). Stimuli were presented via a monaural speaker in a quiet room. Recording procedures for the natural speaker, provisions for tape equivalency between speech sources and presentation levels were not reported. Subjects were asked to verbally report what they heard and were given the option of a second trial for each test item. The sentence materials were presented first to all subjects as they were judged by the experimenters to "require more listening effort" (p.122).

At the single word recognition level significant differences were noted between speech sources, but no differences were noted across age groups. The results also indicated significant differences between speech source for word recognition at the sentence level. In addition, significant differences were seen as a function of age group as well as age group-by-source interaction. The authors indicate however, that the speech stimuli were originally designed for an adult population and therefore the s

s

e

ove

sys

thi

usin

diff

pass

tests

of th

H

Dedina

compre

DECtal

multipl

said to

comprehe

take not

that seen

N., & Mir

difference noted between age groups may reflect the overall linguistic complexity of the sentences. The single-word recognition scores proved to be lower than the sentence scores for the synthesized sources, but not for the human speech source. No differences were noted between stimuli generated using English spelling and phonetic coding for either synthesizer.

Listening Comprehension.

Connected discourse has been used as a part of an overall approach to the assessment of speech synthesis systems. The earliest study available for review which used this approach in isolation was conducted by McHugh (1976) using an early VOTRAX text-to-speech system operated at six different stress settings. The stimuli consisted of passages taken from standardized reading comprehension tests. No differences in performance were noted among any of the experimental conditions.

Hersch and Tartaglia (1983, as cited in Manous, Pisoni, Dedina and Nusbaum, 1985) evaluated the effect of rate on comprehension of speech produced by a prototype of the DECtalk. Stimuli were short passages followed by a set of multiple-choice questions. In this case, questions were said to measure both "literal and "inferential" comprehension of the material. Subjects were allowed to take notes if they desired. Comprehension was similar to that seen for time-compressed speech (Fairbanks G., Guttman, N., & Mirun, M.S.,1957) when synthetic speech was produced

at
res

dist
of t
ques
of c
infe
speec

has b
Frank
1986)
as it
requir
transc
and th
VOTRAX
were no

A s
Pisoni,
With tw
The rat
historic
speech p
been fou
such var

at slow or normal rates. Increases in rate, however, resulted in decreases in comprehension.

Schwab, Nusbaum and Pisoni (1985) also used connected discourse as a measure of listening comprehension in a study of the effects of training . Instead of multiple choice questions, a true/false format was used with several levels of comprehension ranging from word recognition to inferences. Findings revealed no differences between human speech and that produced by the VOTRAX Type N' Talk.

Another approach to assessing listening comprehension has been the use of sentence transcription. Jenkins and Franklin (1981, as cited in Pisoni, Manous, and Dedina, 1986) used two groups of subjects, one transcribe a passage as it was presented sentence-by-sentence, and a second required to await completion of the entire passage before transcription. The speech sources used were a human control and the VOTRAX text-to-speech system. The model of the VOTRAX system was not reported. No significant differences were noted between synthetic and human speech.

A sentence comprehension task was used by Manous, Pisoni, Dedina, and Nusbaum (1985) to compare performance with two human speakers and five speech synthesis devices. The rationale for using this procedure is based on the historical use of sentence verification procedures to assess speech processing with human speakers. Reaction times have been found to be slower when systematic changes were made in such variables as grammatical form (Gough, 1965,1966) and prosody (Larkey and Danly, 1983). The authors theorized " the acoustic-phonetic properties of the speech and its speech quality may affect the amount of time needed to complete any stage involved in the comprehension process" (Manous, Pisoni, Dedina and Nusbaum 1985, p. 38).

Subjects were asked to identify three-and six-word sentences as true or false through the use of a key board and then to transcribe the sentence they heard. Data were obtained for response latency, sentence verification accuracy and accuracy of sentence transcription. The speech sources differed significantly for all three dependent measures. A grouping of sources into three categories was noted. These were labeled (1) natural speech (2) high-quality synthetic speech and (3) moderate- to low-quality synthetic speech. This experimental procedure appears more sensitive than other methods employing connected discourse and multiple-choice questions.

This same paradigm was used to measure the comprehension of sentences presented via digitally encoded speech (Pisoni and Dedina, 1986). Three different methods were used to generate the speech: (1) 2.4 kbps linear predictive coding (LPC) , (2) 9.6 kbps time-domain, harmonic scaling-subband coding (TDHS/SBC), and (3) 16 kbps continuously variable slope delta modulation (CVSD). Statistically significant differences were found in performance for all three dependent variables measured between the highest (CVSD) and the lowest (LPC) ranked vocoders.

A thi

measure c

semantic

Dedina (1

equally o

predictab

final wor

to comple

sentences

sentences

transcript

Items ulti

transcript

Data w

transcript

response l

between th

accuracy.

verificati

Significan

for respon

verificati

scores was

sentences.

interactio

suggested

that diffe

A third study employing sentence verification as a measure of listening comprehension explored the effects of semantic predictability on performance. Pisoni, Manous and Dedina (1986) constructed a set of 80 sentences composed equally of true and false items (40 each) and high- and low-predictability (40 each) items. The predictability of the final word was determined by the number of times it was used to complete a sentence. A pool of 200 potential stimulus sentences was used to generate the frequency data. All sentences were controlled for intelligibility by a sentence transcription task using the DECtalk as speech source. Items ultimately retained for study produced no transcription errors.

Data were collected for the three dependent variables of transcription accuracy, sentence verification accuracy, and response latency. As expected, no differences were noted between the human and synthesized voice for transcription accuracy. The only factor to reach significance in sentence verification accuracy was high- versus low- predictability. Significant differences were noted between speech sources for response verification latency but not for response verification accuracy. In addition, a marked difference in scores was seen for the high- versus low- predictability sentences. Further analysis failed to reveal an interaction between voice and predictability. The authors suggested that this provides evidence against the theory that differences among speech synthesizers are solely the of result segmental intelligibility. Some aspect of the acoustic-phonetic input to the listener interferes with processing meaning as opposed to perception of the sentence. This speculation is further supported by Slowiaczek and Pisoni (1982) response times for a lexical naming task between human and synthesized (MITalk) speech sources. Differences between the speech sources remained the same following training in the task over several days. Taken together, these studies indicate that verification response latency is a sensitive measure for comparison of speech synthesis systems.

Summary

Many high-quality speech synthesis systems produce very low error rates. For example, overall error rates on the MRT were between 3.25 and 7% for the top four systems (Table 1.3). Differences among several synthesizers were as small as 2%. This ceiling effect for natural speech and high-quality synthesizers makes meaningful comparison difficult.

Those studies that employed listening comprehension measures using passages and traditional post-testing failed to demonstrate significant differences among speech sources. These measures do not appear sensitive enough to be useful in comparing speech synthesizers. However, several attributes of test materials increase task difficulty, hence sensitivity to speech source effects. These attributes include (1) open-set response format (Pisoni and Koen, 1981; Greene, Logan, and Pisoni,1986), (2) anomalous sentences

(Pi

196

(4)

Man

ver

lat

to

Man

Ded

"goo

rece

fact

spee

197

be

prod

Qual

incl

phon

arti

reso

rate

indi

recog

(Pisoni and Hunnicut, 1980; Greene, Manous, and Pisoni, 1984), (3) the presence of noise (Pisoni and Koen, 1981) and (4) inclusion of high- and low-probability items (Pisoni, Manous and Dedina, 1986). In addition, sentence verification accuracy and sentence verification response latency seem to be response formats particularly sensitive to difference among systems (Slowiaczek, and Pisoni, 1982; Manous, Pisoni, Dedina, and Nusbaum, 1985; Pisoni and Dedini, 1986; Pisoni, Manous and Dedina, 1986).

Speech Quality

Introduction

Speech quality can be described as the overall "goodness" or naturalness of speech produced processed or received by an element in a communication system. Some factors have been shown to contribute to the perception of speech quality such as intelligibility (Weldele and Millin, 1975); many other attributes have yet to be defined. It can be postulated, however, that each component of speech production contributes in varying degrees to perceived quality. The components and their possible contributions include (1) respiration (via alterations in intensity), (2) phonation (via alterations in fundamental frequency), (3) articulation (via precision of phoneme production), (4) resonance (via changes in oral/nasal coupling), and (5) rate. Thus, speech quality represents more than the individual contributions of speech production, word-recognition, intelligibility, discrimination or prosody: is the total i

attributes

unique. I

these char

limited i

availabil

contributi

may preser

both synth

recognitio

<u>Syllables</u>

It has

synthetic s

presence an

(Simpson an

role of voi

Nusbaum, Gr

via earphon

synthetic v

diotic, dic

Subjects we

the test ta

natural or

as an alter

correctly id

more unique

the total impact of these (and other perhaps undefined attributes) of the speech source which combine to make it unique. In the past, mathematical models used to generate these characteristics in synthesized speech have been limited in their ability recreate this richness by the availability of computer memory and by knowledge of contributing factors. However, speech quality measurements may present a more precise method for differentiating among both synthesized and human speech sources than word recognition or comprehension measurements.

Review of the Literature

Syllables

It has been theorized that the distinct quality of synthetic speech may act to alert the listener to its presence and thus facilitate detection and/or reception (Simpson and Williams, 1980). In order to investigate the role of voice distinctiveness and phonetic discriminability Nusbaum, Greenspan and Pisoni (1986) presented CV syllables via earphones to subjects in the presence of natural and synthetic voice distractors. Levels and mode (monaural, diotic, dichotic) of presentation were not reported. Subjects were asked to identify a target syllable spoken by the test talker from a series of 20 presented by either a natural or synthetic talkers. If the quality of speech acts as an altering mechanism, the percent of the syllables correctly identified should be higher when the target has a more unique character. The results indicated lower recognition performance for both speech synthesizers (DECtalk and VOTRAX Type 'N Talk) compared to the human talker on the basis of all three performance measures (percent correct, response time and false alarm rate). These differences occurred regardless of whether the distracting voices were the same, different or mixed in relation to the target voice. The authors concluded that the distinctiveness of the voice is less critical to target detection than intelligibility of the speech.

Sentences and Connected Discourse

It has been suggested that traditional word-recognition tests fail to provide an adequate representation of the ability to understand speech in normal listening situations (Chial and Hayes, 1974; Oyer and Frankman, 1975; Berger,1978). These tests often lack the accuracy and precision required to demonstrate significant differences among signal sources, transmission systems, or listeners. This suggests that the complex interactions among intelligibility, prosody, message content and the listener's knowledge of the language cannot be evaluated using abbreviated stimulus sets and paradigms.

The IEEE Audio and Electroacoustics Group Subcommittee on Subjective Measurements reviewed a variety of procedures for making speech quality measurements with the intent of discovering those which had been successful and which could be used with a variety of signals. (IEEE,1969) Three methods were recommended: (1) the Category-Judgement

Met

Iso

lis

sig

are

Foo

sco

in

app

spe

int

oft

sig

def

deg

imp

by M

(196

choi

subj

addi

Pref

betw

Method, (2) the Relative Preference Method, and (3) the Isopreference Method.

The Category-Judgement Method requires subjects to listen to a standard speech sample and then compare other signal(s) of interest to this standard. Their impressions are categorized according to the adjectives Unsatisfactory, Poor, Fair, Good, and Excellent. The result is a mean score for the signal based on the total number of judgements in each category. The main difficulty with this method appears to be excessive sensitivity to the content of the speech material used (IEEE,1969).

The Relative-Preference Method places the signal of interest along an arbitrary rating scale based upon how often the signal is preferred in comparison to all other signals. The continuum along which the signal is placed is defined by the selected reference signals; therefore, the degree of degradation used with the reference is of utmost importance (IEEE,1969).

The Isopreference Method (ISM) was originally proposed by Muson and Karlin (1962) and later simplified by Rothauser (1968). A test signal (speech) is presented as a forced choice comparison to a reference signal (also speech) subjected to different degrees of degradation via the addition of noise. Noise level is varied until the preference votes of a listening group are equally divided between test and reference signals. Thus, the signal-to-

no
sc

qu
as
wi
la

us
In
19
ge
as
Whe
co
lis
the

as
com
Vor
wer
Met
req
of
and
subj

noise ratio of the reference signal becomes the preference score for test signal.

An alternative to paired-comparison techniques is quality magnitude estimation. The subject is asked to assign a numerical value to test stimuli. This can be done with (e.g., Chial and Daniel, 1977) or without (e.g., Lawson, 1980) a selected reference signal.

Variations of the Relative-Preference method have been used to differentiate among speech synthesis systems. Nye, Ingemann, and Donald (1975, as cited in Logan and Pisoni , 1986) presented subjects with pairs of short passages generated by several different algorithms. Subjects were asked to state a preference for one of the speech sources. When compared to data obtained from a test of listening comprehension based on the passages, the results revealed listeners tended to prefer the algorithms which generated the highest listening comprehension scores.

McHugh (1976) also obtained data on listener preference as it related to performance on a test of listening comprehension. Six different "inflection levels" of the VOTRAX VS 6.0 were used as speech sources. Twelve sentences were recorded for each source and presented in random order. Method of presentation was not reported. Subjects were required to listen to each sentence and rate the "goodness" of the sample on a seven point scale anchored with "good" and "bad". Results revealed wide variations in how the subjects used the scale. A tendency to cluster at the high

middl
order
prefe
score
rank
resul
revea
on li

I
desig
compa
the p
were
2000
earp
whit
two
voic
coll
Pref
deci
sign
comb
most
pair
sign
inte

middle or low end of the scale was noted. Therefore, in order to combine the data it was necessary to rank order the preferences for each subject by obtaining a mean preference score on the speech source for all twelve sentences. Mean rank was then computed for all subjects. A comparison with results obtained on the test of listening comprehension revealed order of preference was similar to rankings based on listening comprehension performance.

Logan and Pisoni (1986) conducted two experiments designed to evaluate listeners' preferences using a paired-comparison paradigm. Stimuli used in one experiment were the Harvard Psychoacoustic sentences. Three speech sources were used, the DECtalk 2.0 (voice type not reported), Prose 2000 V3, and MITalk-79. Sentences were presented under earphones at an SPL of 80 dB in the presence of 50 dB of white noise. Following presentation of the same sentence by two speech sources, subjects were asked to select which voice was most natural sounding "A" or "B". Data were collected for pair-wise preference, response latency for the preference choice, and confidence rating for the preference decision. All differences in preference were found to be significant with the exception of the Prose/MITalk combination. Logan and Pisoni state that in all cases the most intelligible voices were also the most preferred in the pair. Confidence in rating was also statistically significant and found to correspond to rankings by intelligibility.

In a second experiment the same procedures were employed; however, the Phoneme Specific Sentences were used as stimuli. In addition, the Infovox was substituted for the MITalk as one of the speech sources. A similar pattern was noted among speech sources for pair-wise preference. In addition, the same association between preference and intelligibility ranking was seen.

A different approach was used by Nusbaum, Schwab, and Pisoni (1984, as cited in Logan and Pisoni, 1986). A questionnaire was used to determine subjects' subjective preferences for speech generated by the MITalk and the VOTRAX Type 'N Talk. The questionnaire required subjects to make forced-choice judgements between pairs of adjectives (e.g., gentle/harsh, halting/fluent, hard/easy). The synthesizers tended to be rated as having speech that was more harsh, rough and course than natural speech. It is noted, however, that the preselected adjectives may have biased the subjects' responses. Further, the data provided little information about the attributes used to make these judgments.

Summary

A variety of methods have been used to measure the quality of synthesized speech including questionnaires (Nusbaum, Schwab, and Pisoni, 1984), phoneme recognition (Nusbaum Greenspan, and Pisoni, 1986), scaling (McHugh, 1976) and forced choice pair-wise comparisons (Logan and Pisoni, 1986). A persistent problem in the investigation of

speech qua
scrutiny a
to obtain
precise ar
associatic
and Pisoni
does not r
intelligib
Klatt, 198

Evalu

Most s
term recal
in evaluat
which requ
levels of
measures c
speech com
those inde
intelligib
1986).

A disti
feedback be
and role of
formality c
situations,
acknowledge

speech quality is the definition of the concepts under scrutiny and the operationalization of those concepts so as to obtain results which are quantifiable, repeatable, precise and accurate. Researcher have demonstrated an association between intelligibility and quality (e.g., Logan and Pisoni, 1986), however, a high degree of naturalness does not necessarily mean the the speaker will be intelligible (Nixon, Anderson and Moore, 1985 as cited in Klatt, 1987).

Evaluation of Receiver Performance in Complex Tasks

Introduction

Most studies of synthesized speech involve only short term recall or repetition of stimuli. A plausible next step in evaluation of synthesized speech is to devise a task which requires a variety of skills and offers different levels of complexity. There is a "need to develop new measures of sentence comprehension that can be used to study speech communication at processing levels above and beyond those indexed through transcription tasks and forced-choice intelligibility tests" (Pisoni, Manous and Dedina, p. 20, 1986).

A distinguishing characteristic of human conversation is feedback between receiver and sender. The nature, extent and role of such feedback varies with the purpose and formality of the communication paradigm. In some situations, feedback signals sent from receiver to sender acknowledge receipt of a signal or message (or of the

re
ac
se
ca
hy
se
ma
ot
se
me
th
to
re
sy
ab

sy
co
co
re
hy
ma
H

di
pe
19

receiver's confidence about what was received). Such acknowledgment may invite additional communication from the sender or signal turn-taking in conversation. In other cases, feedback may be used to verify a receiver's hypothesis about a signal or message originating with a sender. Here verification of portions of signal or message may enhance the accuracy of information transfer. In still other cases, feedback from a receiver may convey to the sender the idea that a signal was not received or that a message was not understood. Such information may be used by the sender to modify either the signal or message (or both) to optimize information exchange. Modifications that respond to particular characteristics of the communication system (sender, channel, receiver) may be thought as adaptive.

When a receiver uses feedback to minimize known or suspected errors in communication, the result is that of correcting or repairing flawed information exchange. Thus, corrective feedback can be thought of as "communication repair" behavior, This seems to be a natural event among humans and may influence perceptions of naturalness of man-machine interaction, if not also of synthesized speech.

<u>Human Interaction</u>

Early research in human communication demonstrated that different amounts of feedback increase accuracy of performance (Leavitt and Mueller, 1951; Rosenberg and Hall, 1958). Leavitt and Mueller (1951) designed two experiments with the

or absen

confiden

the task

differen

experime

geometric

feedback

visual o

Yes/no s

Free fee

revealed

(both st

accuracy

the time

additiona

Exper

feedback

was to de

a longer

noted bet

confidenc

condition

whereas t

the same

was evalua

significan

with the purpose of determining the effects of the presence or absence of feedback on such variables as accuracy, confidence level of the subjects and time to completion of the task. Classroom instructors were asked to describe different groupings of geometric shapes. In a first experiment, the students were asked to recreate the geometric patterns from oral descriptions. Four different feedback conditions were employed: 1) Zero feedback (no visual or verbal), 2) Visible audience (no verbal), 3) Yes/no student responses (visual and limited verbal), and 4) Free feedback (unlimited visual and verbal). Results revealed consistent increases in accuracy and in subjects' (both students' and instructors') confidence in their accuracy with increased levels of feedback. Conversely, the time required to give the instructions increased when additional feedback was made available.

Experiment two further explored the conditions of no feedback and free feedback. The purpose of this experiments was to determine the effects of feedback on performance over a longer series of trials (4). The same differences were noted between feedback conditions in initial accuracy, confidence and time. However, in the zero feedback condition, accuracy climbed steadily with repeated trials, whereas the feedback condition began and then remained at the same high level. In contrast, when time to completion was evaluated, the zero feedback condition demonstrated no significant change with trial, whereas the free feedback condit

comple

human

accura

systen

even

furthe

requi

the p

amoun

are c

Human

S

speec

resea

perfo

perce

subje

Nusba

1986

Piso

repe

requ

vari

sema

nove

reve

condition showed a steady decline in the time required to complete the task. Thus, interpersonal feedback during human communication increases subject confidence and accuracy, although the presence of an internal feedback system is suggested by an increase in performance observed even in the absence of free interpersonal feedback. It is further noted that adding more feedback increases the time required to complete tasks. The authors suggested, however, the potential exists for a continuing decrease in time and amount of feedback to the point where most misunderstandings are clarified and no feedback is required.

Human-machine Interaction

Similar findings would be expected when synthesized speech is used in place of the human voice. Although no research is available on the feedback during task performance, researchers have demonstrated improved perception of synthesized speech with feedback to the subject about performance during training sessions (Schwab, Nusbaum and Pisoni, 1985; Greenspan, Nusbaum and Pisoni, 1986). In a study conducted by Greenspan, Nusbaum and Pisoni (1986), training took the form of visual and auditory repetition of the item to the subject (whether or not it was requested) following transcription of the stimulus item. A variety of message types (words, meaningful sentences and semantically anomalous sentences) and training stimuli (both novel and repeated) were used. The post-training tests revealed subjects demonstrated significant improvements in

recogniti
untrained
performan

An ob
talkers i
elicit, u
is possi
that pro
those ob
example,
synthesi
options
the kind
person-
listenin
signal
phrasin
cardina

Thi
of a fe
speech
What ty
request
accurac
alterat

recognition of the stimuli regardless of message type. The untrained control group demonstrated no such improvements in performance.

Summary

An obvious difference between human and synthetic talkers is the difficulty of machine-based systems to elicit, understand and use feedback from a listener. Yet it is possible to provide the listener with control options that produce effects (in the synthetic talker) similar to those observed in person-to-person communication. For example, a listener can be given the option to cause a synthesizer to repeat a previous signal. Other command options can be specified strategically in consideration of the kinds of feedback that enhance information transfer in person-to-person communication conducted in difficult listening situations. Such options include changes in signal level (or in signal-to-noise ratio), speaking rate, phrasing (word choice), and speaking mode (e.g., normal vs. cardinal letter names or coded letter names).

This leads to several questions regarding the presence of a feedback option in an ongoing task in which synthesized speech is used. How frequently would feedback be employed? What types of alterations in the message would the listener request? What are the effects of a feedback option on accuracy of task performance? Do patterns of requested alterations vary with listener experiences. Can answers to these que

the intel

Studi

have been

comprehen

identifi

useful i

The

Shannon-

1.4). F

produced

data car

systems

Part

and cor

intelli

three e

perform

perform

contain

The fol

data:

(1)

these questions promote further analysis of differences in the intelligibility of speech synthesizers?

Purpose of the Study

Studies of the understandability of speech synthesizers have been based primarily on word-recognition and listening comprehension tasks. Review of the available research has identified aspects of those procedures which may prove useful in further evaluations of synthesized speech.

The study was devised to parallel the components of the Shannon-Weaver model of communication systems (see Figure 1.4). Part one entailed an acoustic analysis of signals produced by speech synthesizers. It was theorized that such data can be used to predict performance of speech synthesis systems.

Part two evaluated semantic precision by using sentences and connected discourse presented in noise as measures of intelligibility and listening comprehension. Finally, part three evaluated the effect of synthesized speech on receiver performance and the role of feedback upon receiver performance using a task requiring a variety of skills and containing several levels of complexity.

Questions

The following questions were addressed upon analysis of the data:

(1) Do synthetic speech sources differ significantly from each other (and from human talkers) as a function of

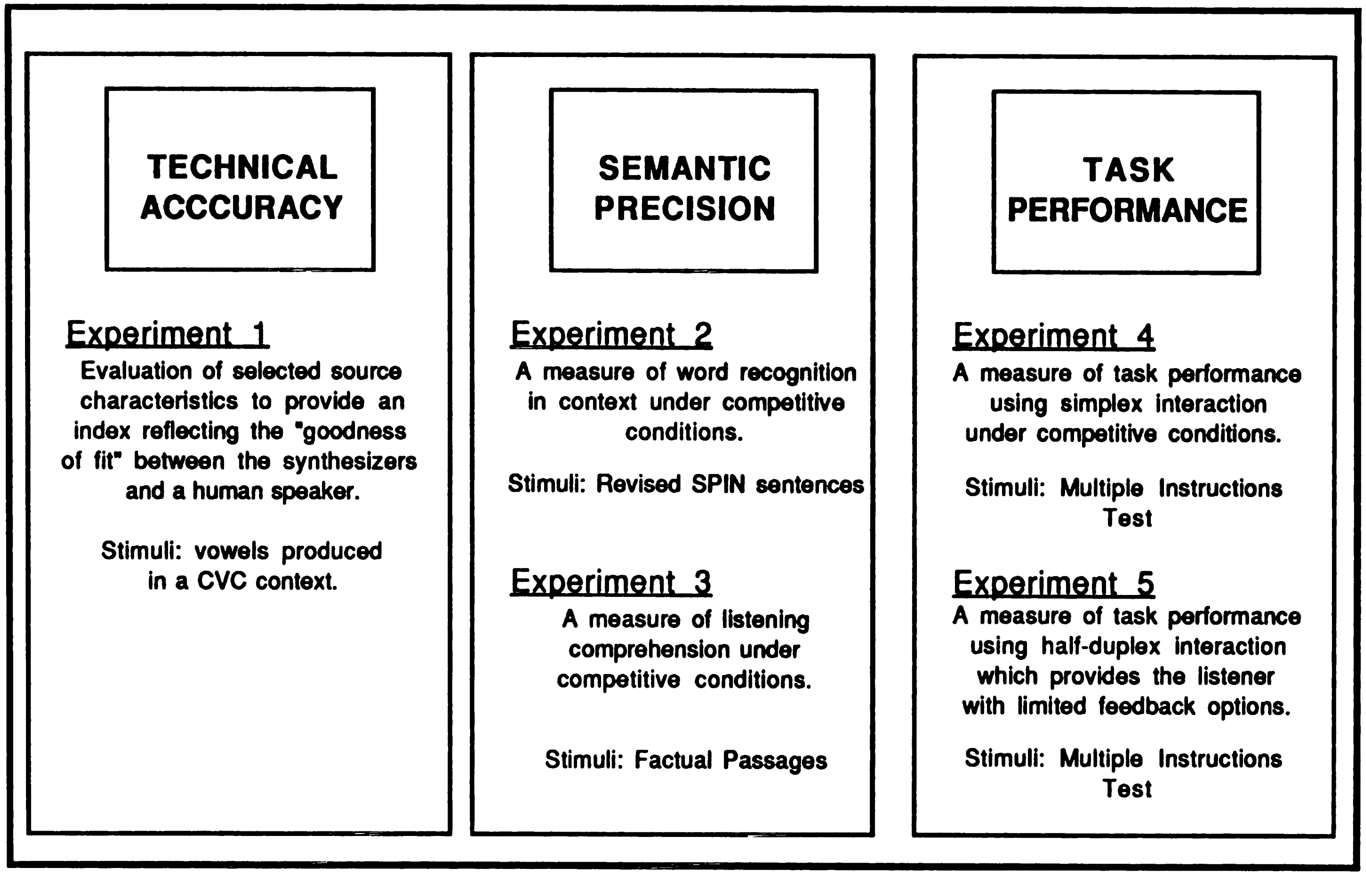


Figure 1.4. Summary of the components of the communication model and the corresponding experiments.

a. acoustic structure of phonemes,
b. word-recognition measured in noise,
c. comprehension measured in noise,
d. complex task performance?

(2) To what extent can b, c, and d (above) be predicted from a?

(3) What is the effect of listener-invoked feedback upon the ability of listeners to accomplish complex tasks directed by speech synthesizers?

Five e

systems en

experiment

synthesis

behavioral

to relate

evaluation

Subjec

ages of 18

language.

(1) He
(r
ar

(2) Ty
ea
(N
be
A

(3) Ac
dB

(4) A
at
th
Ba

(5) No
he
ve

CHAPTER II

METHODS

Introduction

Five experiments were designed to assess communication systems employing speech synthesis devices. The first experiment described signals produced by the speech synthesis systems. The next four experiments described behavioral response to such signals. The overall goal was to relate in the outcomes arising from alternative evaluation methods.

Subjects

Subjects were audiometrically normal adults between the ages of 18 and 35 who claimed English as their native language. Normal hearing was defined as the following:

(1) Hearing Threshold Levels (HTLs) at or below 15 dB (re ANSI S3.5- 1969) for 250, 500, 2000, 3000, 4000, and 6000 Hz in the test ear;

(2) Tympanometric curves characterized by normal middle ear pressure at PEV (between - 100 and +50 mm) (McCandless and Thomas,1974), peak equivalent volume between 0.2 and 1.4 ml and a type classification of A or Ad, unimodal (Jerger, 1970);

(3) Acoustic reflex thresholds (HTL) between 70 and 100 dB at 1000, and 2000 Hz (Northern and Grimes, 1979);

(4) A decrease in acoustic reflex amplitude (for signals at acoustic reflex threshold + 10 dB) not greater than 50% in 10 seconds at 500 and 1000 Hz (Anderson, Barr, and Wedenberg, 1970);

(5) No reported history of otologic surgery, family hearing loss, recent upper respiratory problems, vertigo, tinnitus, or hearing loss.

Subj

synthesi

subjects

subjects

more tha

screenin

Appendix

undergon

Five

synthesi

here:

(1) C
C

(2) I
A

(3) S
I
C

(4) I
W
t

(5) V
(

Syste

circuits

microcomp

on system

software.

accept te

Subjects had no previous experience listening to synthesized speech. Data were collected from twelve subjects for Experiment 2, 3 and 4 and from fourteen subjects for Experiment 5, with no subject participating in more than one experiment. Copies of the audiometric screening form and the informed consent form are provided in Appendix A. Figure 2.1 illustrates experimental events undergone by subjects for Experiments 2 through 5.

Signal Sources

Five devices representing a range of contemporary speech synthesis technology were used in the experiments reported here:

(1) Commodore Amiga (Instrumented using the Commodore Chip # 8364),

(2) Digital Equipment Corporation, DECtalk (model DTC01-AA, "Perfect Paul"), 1983.

(3) Street Electronics, Echo II+, For use with Apple IIe, 1984, (Instrumented using the Texas Instruments Chip #TMS5220CNL).

(4) First Byte, Smooth-Talker (version 2.0), For use with Apple Macintosh (Instrumented via proprietary technology and chip), 1985.

(5) VOTRAX Incorporated, Personal Speech System, 1982, (Instrumented using the VOTRAX Chip #SC01A IC).

Systems 1 and 4 employed proprietary sound generation circuits integrated into commercially packaged microcomputers. Text-to-speech translation was accomplished on system 4 using a rule structure implemented in disk-based software. Systems 2 and 5 were peripheral units designed to accept text from computers or computer terminals via serial

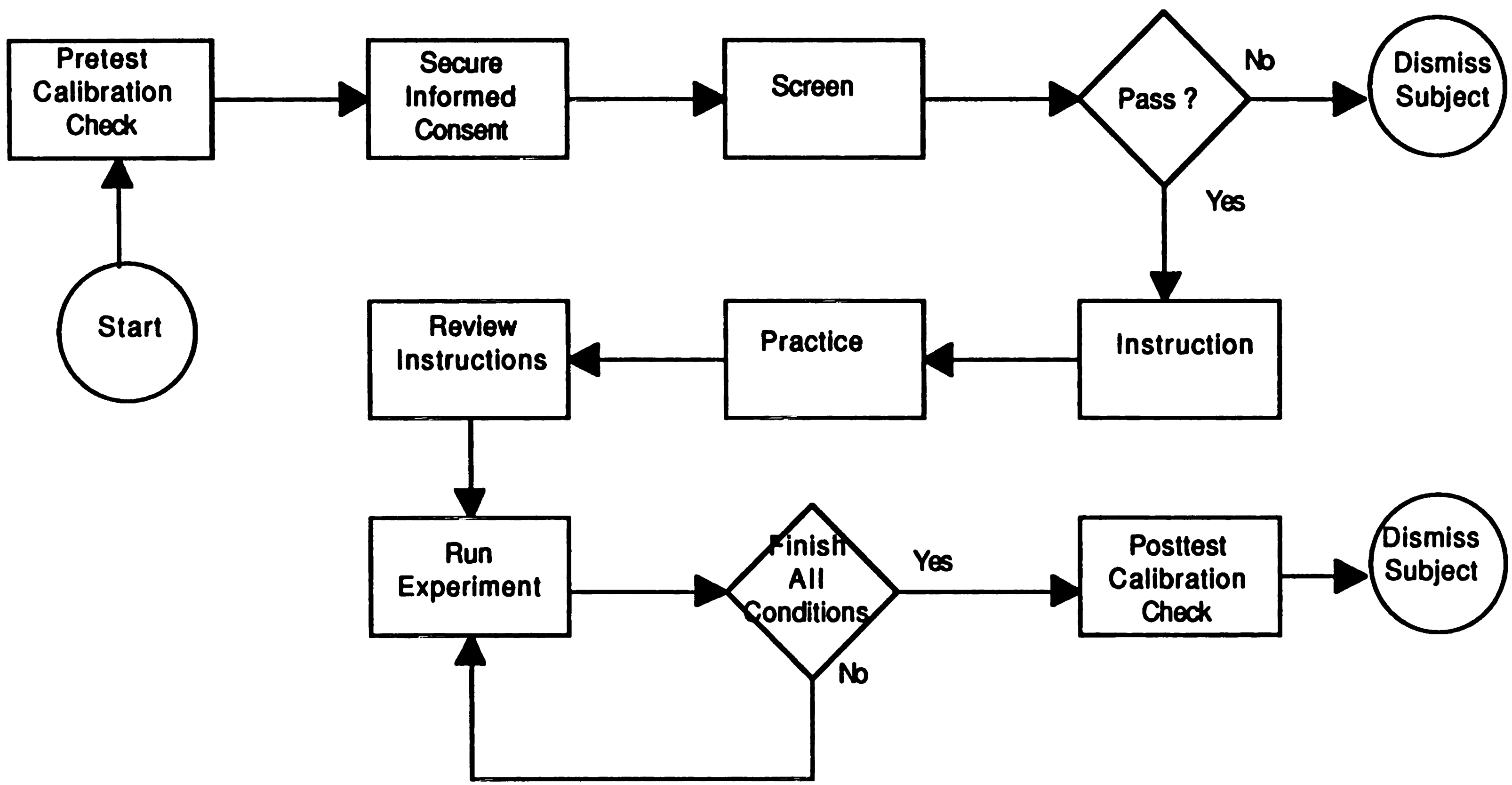


Figure 2.1. Process flow chart of experimental events for experiments 2 through 5.

communi

accompl

memory

was an

periphe

Text-to

distrib

charact

synthes

Unle

were mad

the syn

to a ca

were ge

male sp

for rat

default

automat

efforts

any of

to defa

the sen

loop" we

the rec

communication lines. Text-to-speech translation was accomplished using rule structure implemented by read-only memory (ROM) devices located in the synthesizer. System 3 was an "add-in" circuit board (also available as a peripheral device) containing sound generation circuits. Text-to-speech translation was accomplished via software distributed in disk form. The specific settings and voice characteristics classified as the "default" mode for each synthesizer are summarized in Appendix B.

Synthesizer Configurations

Unless otherwise noted recordings for all experiments were made by on Ampex 632 tape by connecting the output of the synthesizer to an audio mixer (TEAC, Model MB-20), then to a cassette recorder (JVC, Model KD-15). The materials were generated using the text-to-speech systems set for the male speaker or default pitch settings. Default conditions for rate and inflection were employed whenever possible. If default modes did not exist, a mid-scale value was used. If automatic inflection was available, it was employed. No efforts were made to optimize text-to-speech translation by any of the synthesizers. Volume control settings were set to default or mid scale positions. Successive repetitions of the sentence "He tacked the tip tap top of the teep with toop" were output to allow adjustment of the input level of the recorder for a VU peak between -3 and 0 dB.

Exper

acoustica

develop a

reflectin

Stimulus

The s

four back

standard

consisten

was speci

the vowel

CVC nucle

phonetic

Recording

A mal

phonetic

another t

random or

rate and

at each p

measureme

Tape reco

repetitio

the teep

Technical Precision

Experiment 1: Acoustical Analysis

Experiment 1 represented an attempt to describe selected acoustical properties of speech sources. The goal was to develop a single number index for each speech synthesizer reflecting a goodness of fit to human speech.

Methods

Stimulus Materials

The stimuli consisted of three front vowels [i,ɪ,ɛ,æ], four back vowels [u,o,ɔ,ɑ], and one central vowels [ʌ] of standard American English. These vowels were produced in a consistent CVC nucleus environment where the consonant [p] was specified to simplify spectrographic identification of the vowel. As suggested by House and Fairbanks (1953), the CVC nucleus was preceded by [hu] to minimize the effects of phonetic context on test syllables.

Recording Methods

A male talker of General American dialect recorded phonetic stimuli using the sentence "Say the word who pVp another time". Each vowel was presented to the speaker in a random order and spoken at "habitual" inflection, pitch, rate and linguistic emphasis. Three trials of the word list at each pitch level were conducted with the average measurements of those trials used for analysis of data. Tape recorder input level was established using successive repetitions of the sentence "He tacked the tip tap top of the teep with toop" for a VU range of -3 to 0 dB. The recordin

(Electro

cm.

<u>Experime</u>

<u>Spec</u>

properti

fundamen

through

5500 con

DSP Sona

analysis

used for

while ch

fundamen

gated to

Cursors

by audit

selectio

filter (

identifi

visual g

as requi

displaye

Fundament

the peak-

The recip

fundament

recording were made in a sound chamber using a microphone (Electro-Voice, Model RE-15) to mouth distance of 20 to 30 cm.

Experimental Procedures and Apparatus

Spectrographic Analysis. To describe the acoustical properties of the signal sources, measurements of rate, fundamental frequency and center formant frequencies (f_1 through f_3) were performed using a Kay DSP Sona-Graph model 5500 connected to a Marantz Model 5020 cassette player. The DSP Sonagraph allows for dual-channel, real-time display and analysis of acoustic signals. Therefore, channel one was used for a spectographic display and format measurements, while channel two was used for display waveform display and fundamental frequency measurements. Recorded stimuli were gated to isolate the CVC sequence from the carrier phrase. Cursors were used to isolate the target syllables, followed by auditory playback of the segment to verify the selections. Spectrograms were produced using a broadband filter (300 Hz) on a time axis of 1.0 second. Formants were identified by visual inspection using the black on white visual graphics. Various color coding options were selected as required to enhance visualization. The DSP automatically displayed the frequency identified by the cursor. Fundamental frequency measurements were made by measuring the peak-to-peak time intervals on the wave form display. The reciprocal of these time intervals estimated the fundamental frequency of CVC vowels.

Expe

designed

goal was

under co

<u>Stimulus</u>

Word

Speech Int

original g

listening

materials

Elliott, 1

Dubno and

Morgan, D.

demonstrat

consists c

contains 2

words. Te

competing

others (19

based upon

babble rat

level of 5

are scored

Semantic Precision

Experiment 2: Word Recognition.

Experiment 2 represents the first of three studies designed to evaluate selected aspects of reception. The goal was to obtain a measure of word recognition in context under competing message conditions.

Method

Stimulus Materials

Word recognition was assessed using the Revised Test of Speech Intelligibility in Noise (SPIN) (Bilger, 1983). The original goal of this test was to provide a measure of listening abilities in everyday situations. Evaluation of the materials with both normal-hearing (Kalikow, Stevens and Elliott, 1977) and impaired-hearing subjects (Dirks, Kamm, Dubno and Velde, 1981; Dubno, J. R., Dirks, D. D., and Morgan, D. E.,1984; Bilger and others, 1984) have demonstrated usefulness of the test. Bilger's revision consists of eight recorded lists of 50 sentences. Each list contains 25 high-predictability and 25 low-predictability key words. Test materials are presented together with a competing signal (twelve-talker voice-babble). Bilger and others (1984) offer norms for normal and impaired listeners based upon presentation of test sentences at a speech-to-babble ratio of +8 dB where the sentences are adjusted to a level of 50 dB above threshold for that signal. Responses are scored on the basis of the number of key words correctly identified

Appendix B

SPIN

recognitio

as well as

predictabi

(means sta

each of th

speech sou

three perc

items (25

percent- c

transforme

<u>Recording</u>

To ma

stimuli an

by differe

were measu

between th

segmental

were under

a Larson-D

(Model 800

equipment

placed at t

related to

identified. Copies of the SPIN scoring forms are given in Appendix B.

SPIN test results were indexed by percent-correct word-recognition scores computed separately for total test lists, as well as for low-predictability key word and high-predictability key word subtests. Descriptive statistics (means standard deviations and ranges) were calculated for each of these dependent variables, and for each of the six speech sources employed in this experiment. Because the three percent variables were based on different numbers of items (25 for subtests; 50 for full lists), and because percent- correct indices often are skewed, these data were transformed for further analysis.

Recording methods

To maintain a consistent relationship between the speech stimuli and babble tape as well as among stimulus recordings by different speech sources, equivalent sound levels (L_{eq}) were measured. L_{eq} measurements also resolved differences between the synthesizers in terms of relative levels among segmental unite (phonemes). Duration and L_{eq} measurements were undertaken using a TDH-39 earphone, a 6-cc Coupler, and a Larson-Davis Laboratories integrating sound level meter (Model 800B). A block diagram of the arrangement of the equipment is shown in Figure 2.2. Level calibration tones placed at the beginning of each experimental tape were related to the program material such that the calibration

Figur
place

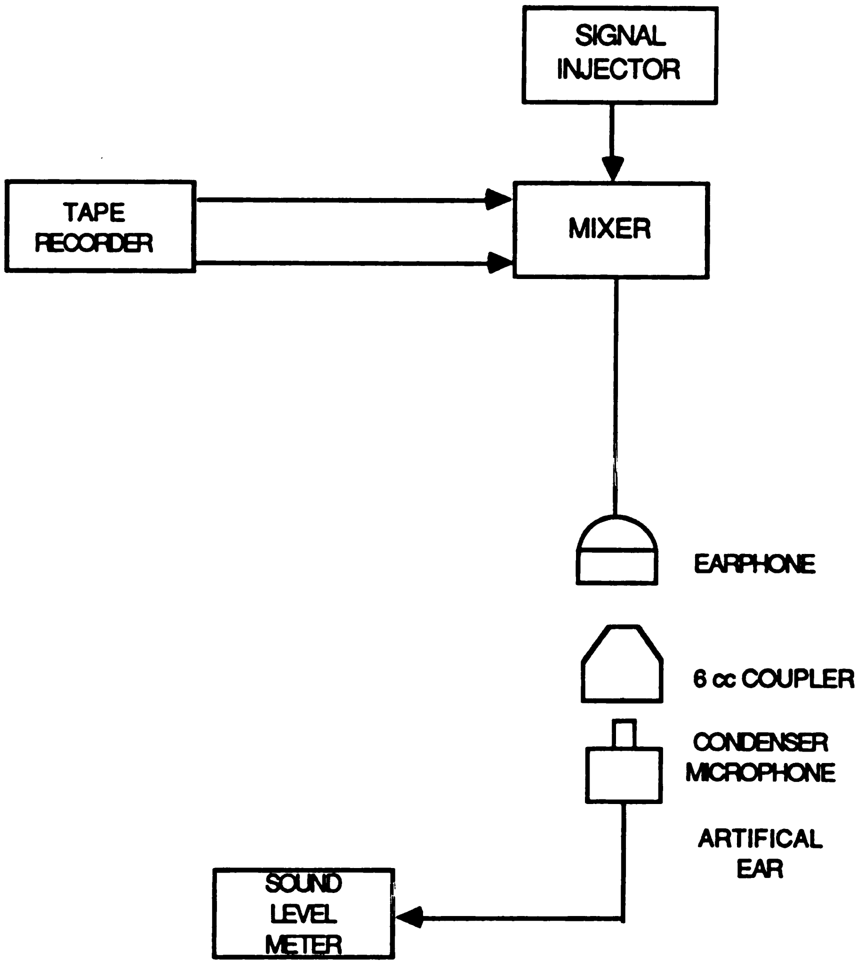


Figure 2.2. Apparatus used for making Leq measurements and placement of 1000 Hz calibration tones on experimental tapes.

tone had

associate

SPI

in Experi

provided

deviates

the low-p

reliabili

measureme

as exper

Experime

Des

experime

subjects

Dimensio

presente

4I/AR cu

frequenc

was cove

reproduc

players

respecti

Stadler,

mixed fo

Cal

experim

calibra

tone had an L_{eq} nominally 6 dB more than the L_{eq} of the associated material.

SPIN lists 1, 3, 5, 6, 7, and eight were used as stimuli in Experiment 2. Selection of these lists was based on data provided by Bilger (1983). He demonstrated that list 2 deviates form other lists in reliability of raw scores for the low-predictability subtest. List 4 has the lowest reliability (r=0.927) and the highest standard error of measurement (7.72%). Therefore, lists 2 and 4 were not used as experimental stimuli.

Experimental Apparatus

Description. A block diagram of the arrangement of the experimental apparatus is provided in Figure 2.3. Individual subjects were seated in a double walled sound suite (IAC, Dimensions 2.54 m x 2.74 m x 1.98 m). Speech stimuli were presented monaurally under earphones (TDH-39 mounted on a MX-41/AR cushion) to the better ear as determined via a four frequency average (500, 1K, 2K and 4K Hz). The poorer ear was covered by a dummy earphone and cushion. SPIN lists were reproduced and voice-babble were recorded with cassette players (Marantz, Model 5020 and JVC, Model KD-15 respectively), and routed to a speech audiometer (Grason-Stadler, Model 162), where the signal and babble tracks were mixed for output to the earphone.

Calibration of listening apparatus. In those experiments involving taped presentation to listeners, a calibration selector switch allowed the examiner to monitor

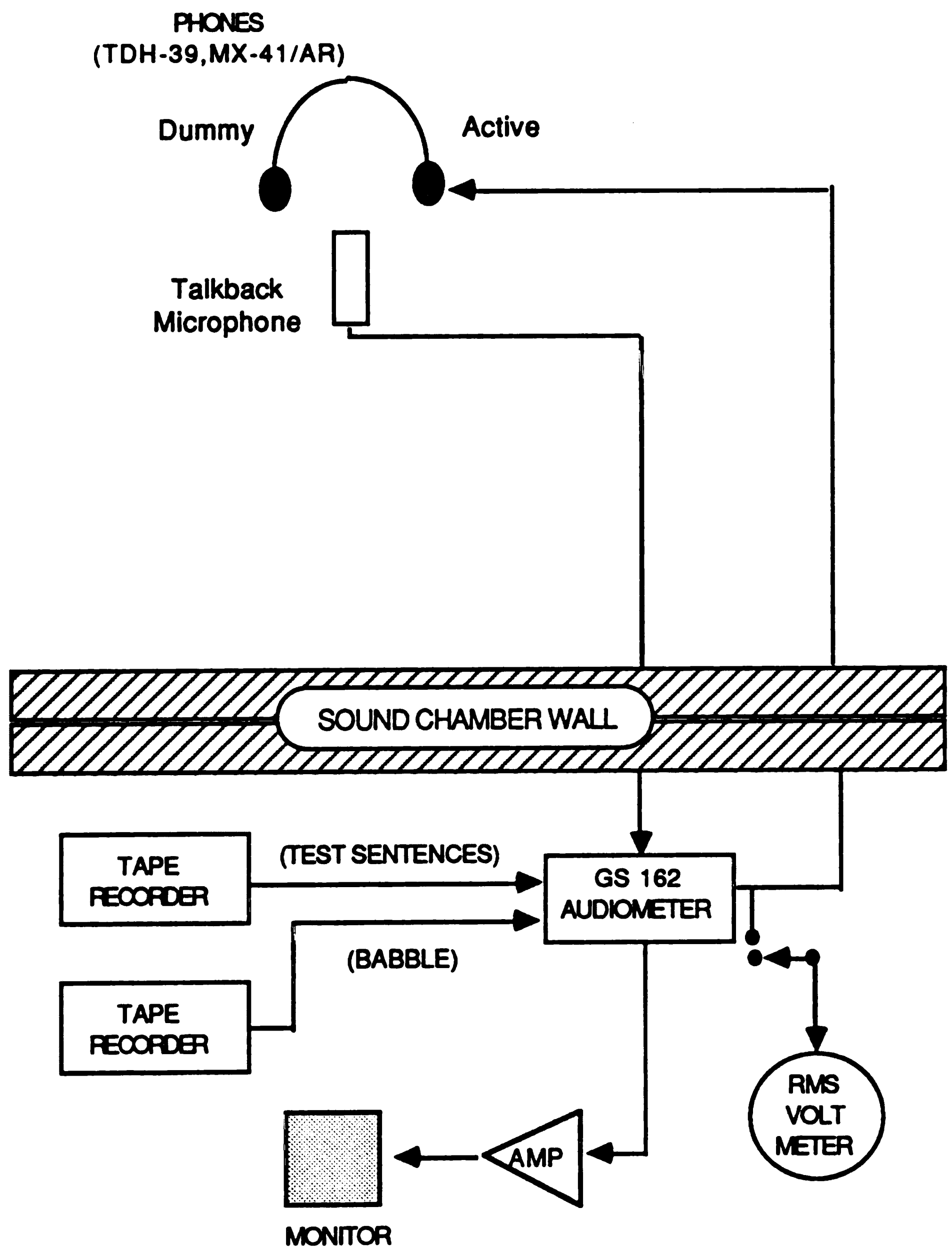


Figure 2.3. Block diagram of experimental apparatus for Experiment 2.

a taped

channel

(Bruel

each SP

a heari

list an

and pos

respons

In

speech

within

freque

found

Th

were

perfo

synth

condu

used

pilot

perfo

qual

the

stud

DECta

spee

a taped calibration tone routed through either audiometer channel. These signals were monitored with a true RMS VTVM (Bruel and Kjaer Model 2409) as follows. Before and after each SPIN list was presented, the audiometer was adjusted to a hearing level (HL) of 70 dB. The VTVM levels for test list and babble calibration tones were prerecorded. If pre- and post-test level checks were within ±1 dB, subject responses were accepted.

In addition to the within-session calibration, the speech audiometer and earphones were checked and found to be within tolerances specified by ANSI S 3.6-1969. The frequency response curve for the TDH-39 earphones can be found in Appendix C.

The signal-to-babble ratios recommended by Bilger (1983) were based upon signals recorded by a human. Because performance-intensity functions may differ for human and synthesized sources (Chial, 1973), a pilot study was conducted to identify a speech-to-babble ratio that could be used for all of the sources studied here. The goal of the pilot study was to find a compromise S/B for which listeners performance with a high quality source (DECtalk) and a low-quality source (Echo II+) fell within the linear portion of the performance intensity function. Details of the pilot study are given in Appendix D. Each subject listened to the DECtalk and Echo II + synthesizers at the four different speech-to-babble ratios: +8, +4, 0, and -4. On the basis of this in

for use

Proced

Aft

the sou

was mea

to be 5

present

SPIN 1

Sub

se
fr
wi
Fo
co
ha
pl
pe
co
yo
g

te
di
af

ar
sy
Le

Ea

the ex

admini

S/B).

this information a signal-to-babble ratio of +8 was selected for use in the Experiment 2.

Procedures

After auditory screening, the subjects were seated in the sound room. A monaural voice-babble detection threshold was measured. The SPIN list presentation level was adjusted to be 50 dB above this threshold and the voice-babble presentation level was adjusted to be -8 dB relative to the SPIN list presentation level (Bilger,1983).

Subjects were instructed as follows:

> "This is an experiment in which you will hear several sets of sentences. The sentences will come from the earphone on your _______ ear. Your job will be to repeat the last word of each sentence. For example, if your hear "Mrs. Smith did not consider the door," then say "door." It will be hard to hear the sentences because they will be played in the presence of background noise of many people talking at the same time. The noise will come from the same earphone as the sentences. If your are not sure of the last word, feel free to guess.
>
> We will use 6 tests, with 50 sentences. Each test takes a few minutes and will be presented by a different talker. There will be a short break after each test.
>
> Once again, your are to listen to the sentence and repeat the last word you hear. If you are not sure of the word please guess. Any questions? Let's try a practice list."

Earphones and the talkback microphone were positioned by the experimenter and a 50 sentence practice list was administered in the presence of the voice-babble- noise (+8 S/B). The practice list was divided into two parts. The first 2

talker

experie

consist

speech

to the

experi

Ex

design

goal w

functi

Stimul

Th

corres

McCal

(1974

readal

score

score

refer

obtai

factu

Junior

use i

used a

first 25 sentences consisted of items presented by the male talker as recommended by Bilger (1983), thus providing experience with the task. The second set of 25 sentences consisted of five sentences produced by each of the five speech synthesizers. This was intended to provide exposure to the different talkers to be used during the test. Total experimental time was approximately 1.5 hours per subject.

Listening Comprehension

Experiment 3 was the second in the series of experiments designed to evaluate selected aspects of reception. The goal was to obtain a measure of listening comprehension as a function of talker.

Method

Stimulus Materials

The stimulus materials consisted of six passages and corresponding multiple choice tests selected from the McCall-Crabs Standard Test Lessons in Reading by Connors (1974). He based his selection was based on an index of readability (discussed below) and on test "difference" scores. Difference scores were calculated by subtracting scores obtained when the tests were administered without reference to the passages ('test only" scores) from scores obtained when the passages were administered. Additional factual passages selected from texts designed for use by junior high and high school students (Chial, 1973) and for use in perceptual research (Cox and McDaniel, 1984) were used as practice stimuli. Phonetic content was not a criteri

materia

Two

to dete

Gunning

(Gross,

represe

compreh

5 and 7

Analysi

(Gross,

These c

Connors

<u>Recordi</u>

Rec

describ

establi

signal,

materia

greater

<u>Experim</u>

<u>Des</u>

experim

environ

method

2.

criteria for selection of material. Copies of these materials are given in Appendix E.

Two measures were applied to the experimental passages to determine levels of readability across passages; Gunning's Fog Index and Flesch's Index of readability (Gross, 1986). Both indices provide values said to represent the number of years of schooling required to comprehend the material. A grade level criterion of between 5 and 7 was used for selection of the experimental passages. Analysis of the passages were performed by means of Thunder (Gross,1986), a program designed for the Macintosh computer. These data, as well as those originally calculated by Connors (1974), are provided in Appendix E.

Recording Methods

Recording procedures were the same as those previously described. Equivalent sound levels (L_{eq}) were used to establish equal levels for taped stimuli and the competing signal. Level calibration tones were related to the program material such that the calibration tone had an L_{eq} of 6 dB greater than the L_{eq} of the associated material.

Experimental Apparatus

Description. A block diagram of the arrangement of the experimental apparatus is provided in Figure 2.4. The test environment, mode and level of signal presentation, and method of level calibration were the same as in Experiment 2.

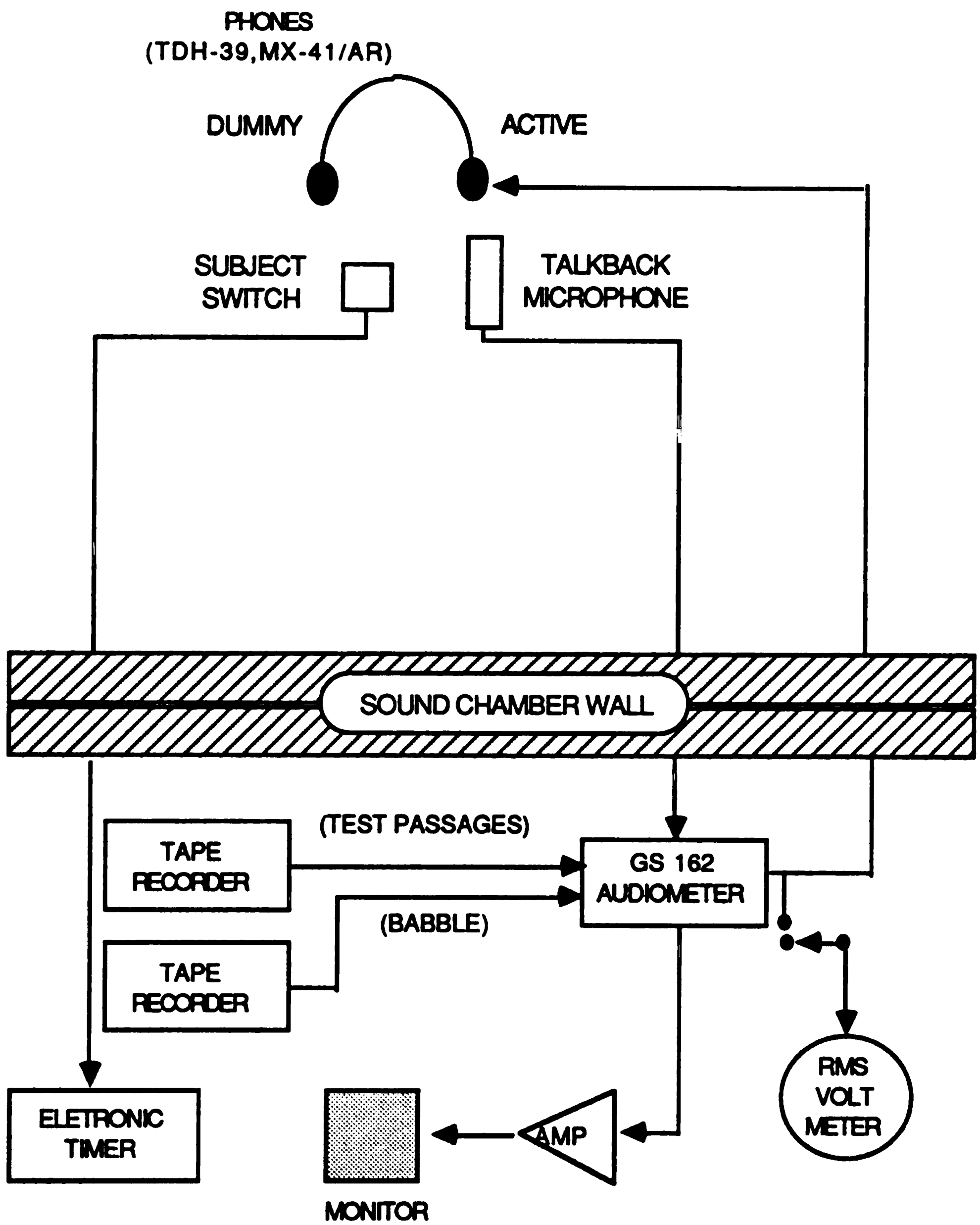


Figure 2.4. Block diagram of Experimental apparatus for Experiment 3.

In

time, t

was con

output

(Coulbo

precis

Hz and

The ti

onset

Proced

Ea

was se

subjec

which

conte

n
t
c
b
t
t
n

b
t
q
a

In order to measure multiple-choice test completion time, the subject was also given a response button. This was connected to a switch input (Coulbourn S22.02) the output of which was connected to a RS/T flip flop (Coulbourn, S41-02). The Flip Flop was connected to a precision time base generator(Coulbourn S51-1-) set to 100 Hz and finally to an electronic timer (Coulbourn R11-25). The timer was reset and activated by the experimenter at the onset of each test using a switch module (Coulbourn S96-03).

Procedures

Each of 12 subjects who satisfied audiometric criteria was seated at a desk and provided with a pencil. The subject was instructed to listen to each passage, after which they were given a multiple-choice test regarding the content. The following instructions were given orally:

> You are going to hear a recordings of passages made by several different talkers. You will hear the speaker in only one ear. The passages will come from the earphone on your _____ ear. It will be hard to hear them because they will be played in the presence of a background noise of many people talking at the same time. I want you to ignore the noise completely. Pay attention only to the voice.
>
> Once you have listened to the passage you will be given a short test regarding the content. The test will consist of eight or nine multiple choice questions. Do not leave a question blank. If you are not sure of an answer please guess. Keep in mind you will be scored on the time it takes you to complete the test as well as the number correct so work as quickly as you can. I will tell you when to begin and when you are done press this button to stop the timer.
>
> There are 6 passages each lasting approximately 30 seconds. At the end of each multiple choice

test there will be a short break. Remember, your job is to listen to the passage and then answer the test questions about the passage as quickly as you can.

Any questions? Let's try a practice set.

Practice passage one (recorded using a female voice) and corresponding test were administered to familiarize the subject with the test taking process. A second practice passage without a test was presented to familiarize the subject with the talker for that list. The same familiarization passage was used prior to each voice. Experimental passages were presented monaurally at a sensation level of 50 dB (re: voice-babble threshold) and at a S/B of +8 dB. Following presentation of an experimental passage the subject was directed to select a folder containing a copy of the associated multiple choice test. The experimenter told the subject to begin and simultaneously started the timer. The subject pressed the timer response button when the test was completed. Order of passages and talkers were randomized among subjects. Approximately one hour was required for each subject to complete Experiment 3.

Task performance

Experiment 4: Oral Instructions

Experiment 5 was the first of two studies evaluating selected aspects of receiver performance. The goal was to obtain a measure of listener performance as a function of talker on a relatively complex task.

Method

Stimulus Material Preparation

Subjects completed an adaptation of the Oral Directions subtest of the Detroit Tests of Learning Aptitude (Hammill, 1985). The original subtest is intended to measure "listening comprehension, spatial relations, manual dexterity, short term memory, and attention" (Hammill, 1985, p. 56). The relative contribution of these abilities to test performance are not known. The subject is given a series of commands to be carried out using pencil and paper, e.g. putting an X in a square, the letter F in a triangle and a numeral 4 in a circle. Thus, each command may be described as containing a minimum of an action-patient combination with an adverb of place. Here action denotes the presence of a verb expressing an activity or movement that can be seen or heard. Patient is defined as the receiver of the effect of a process or action (Heidinger, 1984).

In the present experiment, it was desirable to evaluate the effects of message complexity upon subject performance. In addition, several versions of the test were required for practice and for use with each speech source. Therefore, a modification of the Oral Directions Test was devised. This modification called the Multiple Instructions Test (MIT) and invokes five levels (A through D) of increasing length and complexity. As Figure 2.5 and Table 2.1 illustrate, Level A contained two commands, Level B had three commands, Level

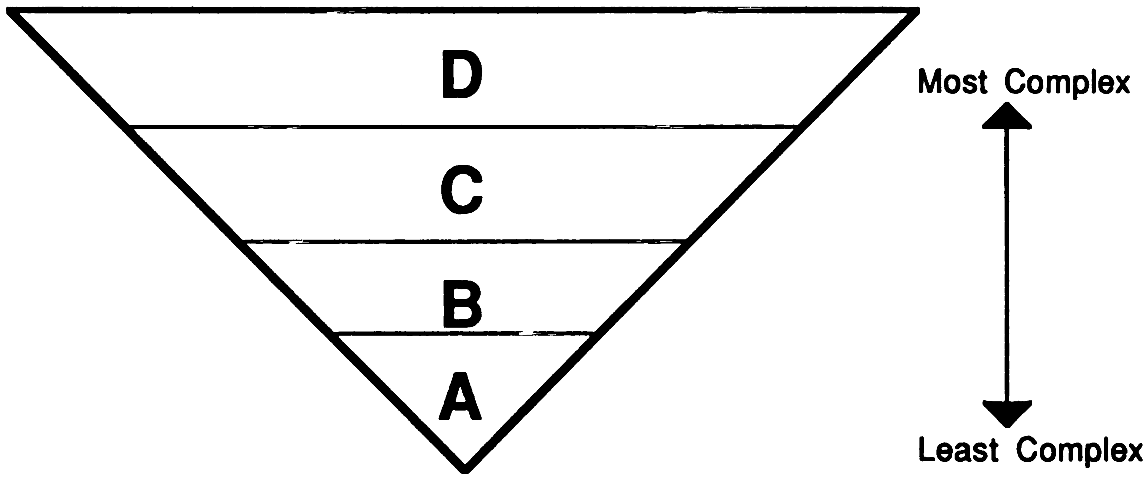


Figure 2.5. Hierarchy of complexity levels used in the Multiple Instructions Test

	Complexity Level of Command A	B	C	D
Number of Commands	2	3	4	5
Number of Words Per Item	9 - 12	15 - 18	21 - 24	27 - 30
Gunning's Fog Index	4 - 5	7 - 9	10 - 12	12 - 15
Flesch's Index	182 - 189	171 - 179	155 - 166	141 - 151
Action and Patient	✓	✓	✓	✓
Adverb- Place	✓	✓	✓	✓
Adjectives				
Size		✓	✓	✓
Ordinal			✓	✓
Color of Object			✓	✓
Restriction			✓	✓

Table 2.1. Permitted content options for each level of command of the MIT.

C had four commands, and Level D had five commands imbedded within each item. Levels B, C and D also contained adjectives (order, color, size). Levels C and D contained restrictions such as " The line may not touch any other object" in place of an action-patient command. For the present purpose a restriction was defined as a statement following a command or series of commands that places additional constraints on the action to be taken. A total of six items were presented at each level. Table 2.1 summarizes the characteristics assigned to each level of command. A list of the actions, patients, adverbs and adjectives, as well as copies of the stimuli are provided in Appendix F.

Subject performance was measured in terms of (1) time-to- completion of total task, (2) time-to-completion per item, (3) percent of items correctly executed at each level of command, and (4) total percent correct for all items. Determination of correct command execution was based on a predetermined set of scoring criteria for each item. Examples of an item, the scoring criteria for each command and sample responses are shown in Table 2.2.

Recording Methods. The basic recording procedures were the same as those previously described. L_{eq} measurements of recorded commands were taken using a Larson-Davis Laboratories sound level meter (Model 800B). Calibration tones placed at the beginning of each experimental tape were related to the program material such that the calibration

Sam

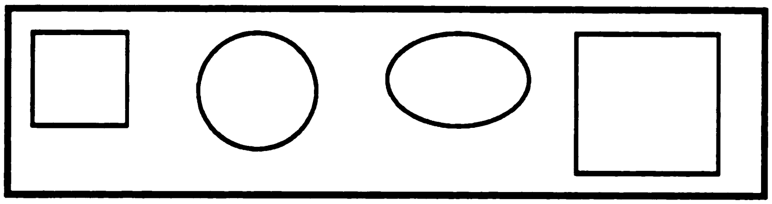
Sam

Sa

dra
tou

Sa

Ta

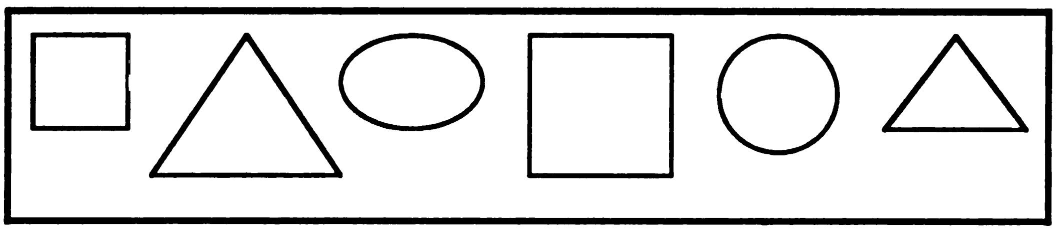


Sample Level A Command:

Put a "B" in the circle and circle the small square.

Sample Scoring Criteria:

1. Is there a "B" in the circle?
2. Is there a circle around the small square?



Sample Level C Command:

Put an "X" in the red square, an "X" in the second triangle and draw a line from the small square to the oval. The line may not touch any other shape.

Sample criteria for scoring:

1. Is there an "X" in the red square?
2. Is there an "X" in the second triangle?
3. Is there a line from the third square to the circle?
4. Does it touch any other shape?

Table 2.2 Sample commands and scoring criteria for the MIT.

tone had an L_{eq} of 6 dB more than the L_{eq} of the associated material.

Additional Screening Procedures.

In addition to the audiometric screening procedures previously described, the subjects for this experiment were asked to undergo screening for color blindness and appropriate color nomenclature. The purpose of this test was to rule out those subjects unable to respond effectively to items containing color modifiers. This was be accomplished through the use of Dvorine Pseudo-Isochromatic Plates (Dvorine, 1953). The subjects were asked to provide the name for a group of saturated colors (red, brown, purple, yellow, blue, green, gray, and orange). Part two called for the identification of the digits on 15 plates made of eight different color combinations to rule out color blindness.

Experimental Apparatus

Description. Figure 2.6 illustrates the equipment used in Experiment 4. Test items and voice babble was reproduced (Marantz, Model 5020 and JVC, Model KD-15 cassette recorders, respectively) , and routed to the speech audiometer (Grason-Stadler, Model 162), where the signal and voice babble tracks were mixed for output through a monaural earphone. The subject was provided with a response button connected to an electronic timer described in Experiment 3. The subjects were asked to press the button when they finished each item. Response latency was timed from the

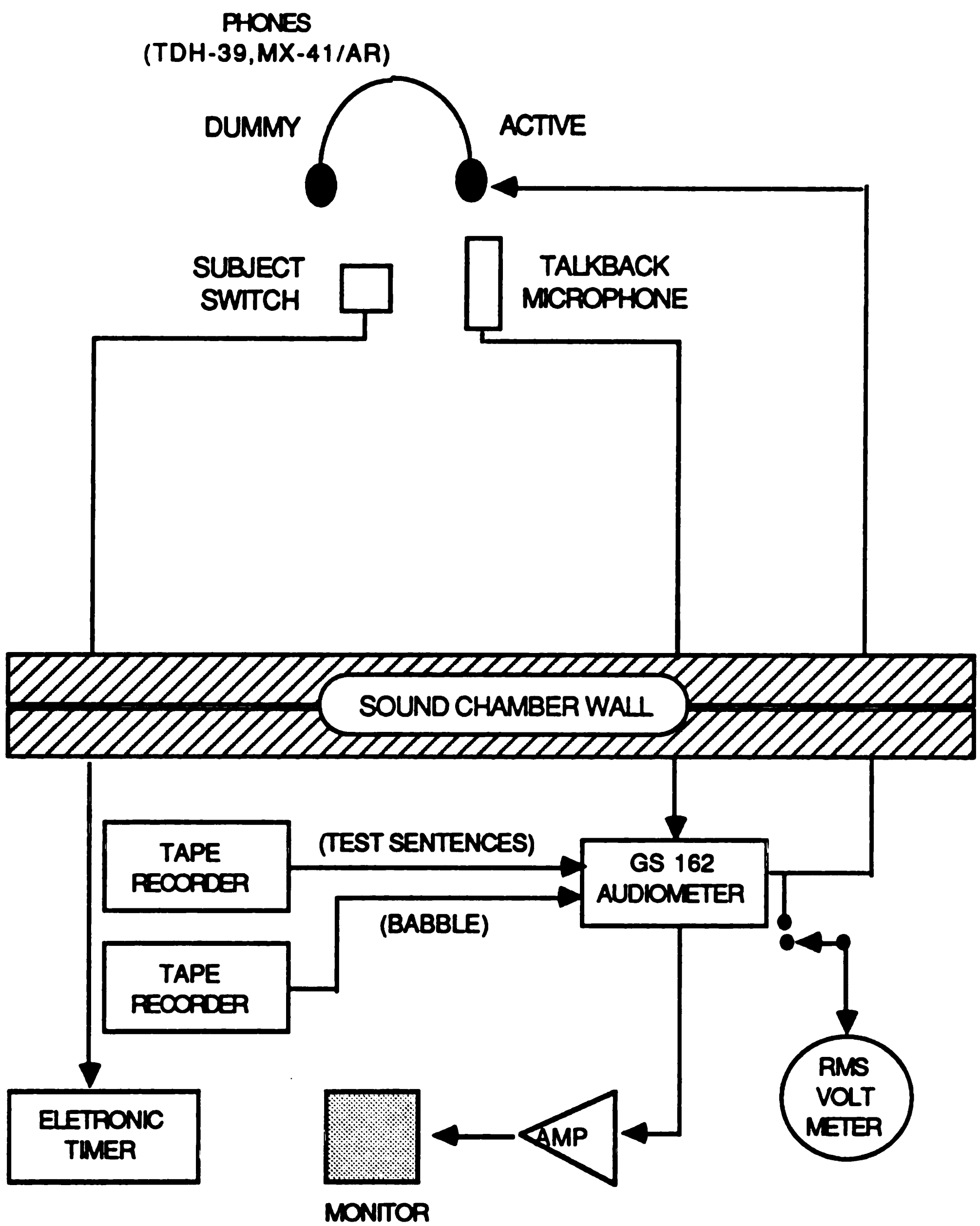


Figure 2.6. Block diagram of experimental apparatus for Experiment 4.

onset of an alerting phrase ("Do it now","Begin now" or "Start").

Procedures

The subjects were provided with the response books, a pen and the response button. They were instructed orally as follows:

> You will be hearing a set of directions to be carried out with the materials in front of you. Here are samples of the pictures and the colors you will be seeing and the colors used to describe them. Take a moment to look them over. Please tell me now if you think you will have any difficulty identifying any of these colors, objects or shapes.
>
> This page demonstrates how the picture will be arranged. A sample instruction might be "Number one. Put a P in the diamond and circle the square." Please listen carefully to the directions given for each item, waiting to begin until after you have heard the command to do so. Immediately upon completion of the item push the response button provided. Respond as quickly and as accurately as possible for each one. Your score is based on the time it takes you to complete it correctly.
>
> The directions you will hear are made by several different talkers. You will hear the speaker in only one ear. Along with the voice you will hear a noise of a group of people talking at the same time. I want you to ignore the noise completely. Pay attention only to the voice and carry out the instructions as quickly and as accurately as you can.
>
> Any questions? Let's try a practice test.

The practice list consisted of 24 items with one item presented by each of the six speech sources at each complexity level. This served to provide subjects with experience in both the experimental task and speech sources.

Practice and experimental tests were presented at a level of 50 dB above the subjects threshold for the voice-babble. A signal-to-babble of +10 was be maintained. Order of tests and talker was randomized between subjects. A short break was provided between each test. The data acquisition time for each subject was approximately 2 hours.

Experiment 5: Oral Instructions With Communication Repair

The goal of experiment 5 was to describe the pattern and effect of communication repair options on complex task performance.

Methods

Stimulus Materials and Experimental Apparatus

Description. Subjects were asked to complete the MIT using the VOTRAX Personal Speech Synthesizer (PSS). This device was selected because it could be interfaced with the Apple IIe computer for listener control and because it offered a reasonable number of repair options. In addition, initial analysis of the results of Experiment 2 indicated the VOTRAX PSS ranked third of the five synthesizers. By selecting a mid-range device it was hoped to avoid ceiling effects when repair options were employed. Version 4 of the MIT was selected for use with this experiment. Item presentation was controlled by the computer using a program written for this purpose. All experimental items were presented live, thus allowing the subjects to choose from a limited set of repair options. The calibration phrase "He tacked the tip tap top of the teep with toop" was used to set the VU meter of the audiometer to a range of -3 to 0 dB prior to testing each subject.

Figure 2.7 illustrates the equipment configuration for Experiment 5. Voice-babble was reproduced via cassette recorder and mixed with the MIT items for monaural earphone presentation. MIT stimuli were presented 50 dB SL (re: voice babble detection threshold; A voice-to-babble ratio of +10 was used for all subjects. If a subject selected the "repeat louder" option the S/B returned to +10 dB after that item was presented. Sound level measurements of the stimuli indicated the repeat louder option resulted in an increase of 5 dB. Consequently the S/B was increased to +15 dB when this option was selected. Control of the repair option was performed by the experimenter via keyboard following instructions from the subject.

Procedures

Each subject was provided with response books and a pen. The following instructions were given orally:

> You will be hearing a set of directions to be carried out with the materials in front of you. Here are samples of the pictures and the colors you will be seeing and the colors used to describe them. Take a moment to look them over. Please tell me now if you think you will have any difficulty identifying any of these colors, objects or shapes.
>
> This page demonstrates how the picture will be arranged. A sample instruction might be "Number one. Put a P in the diamond and circle the square." Please listen carefully to the directions given for each item. Your job is to carry them out as accurately as you can. If you wish, you may ask

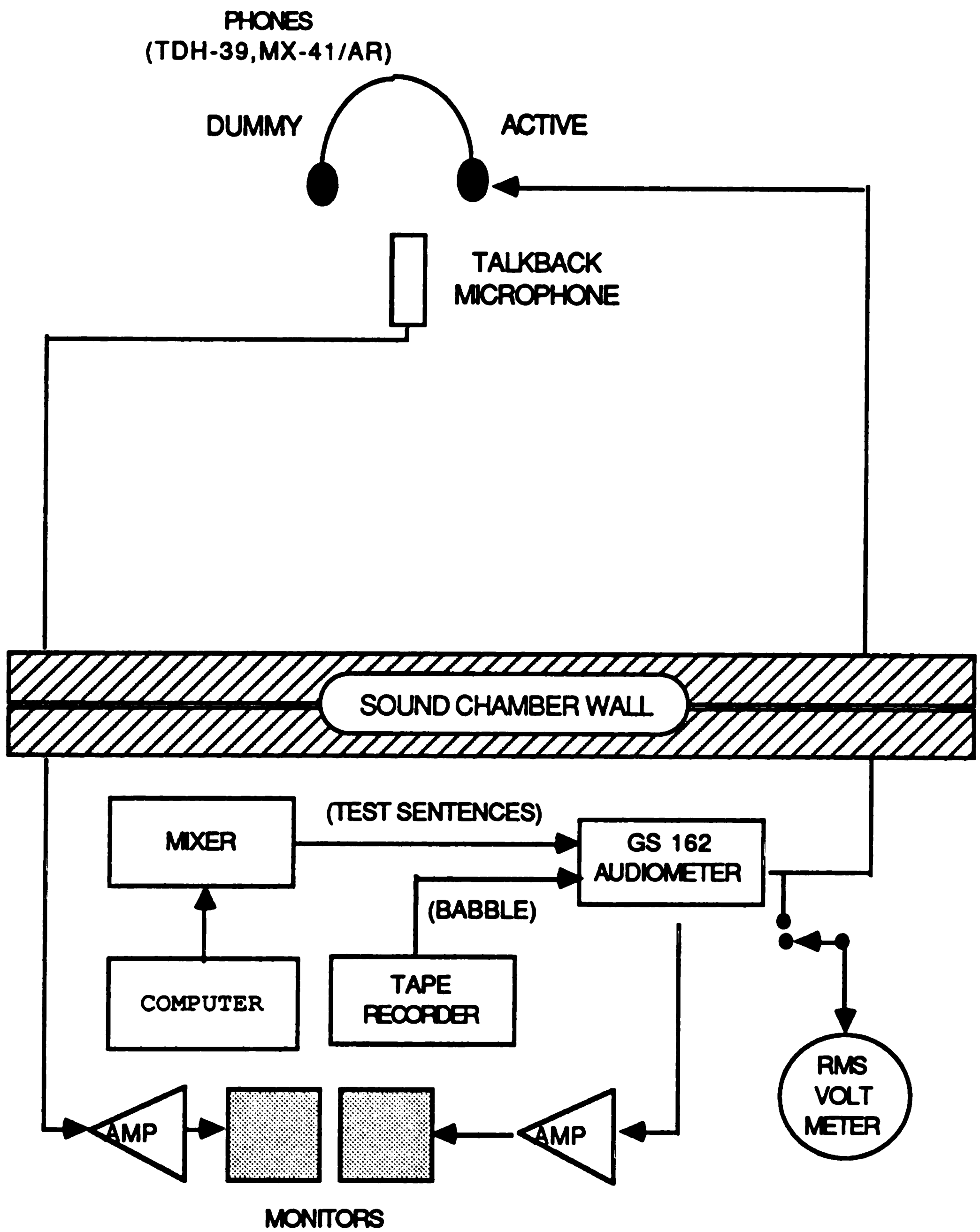


Figure 2.7. Block diagram of experimental apparatus for Experiment 5.

for
fol

(2)
(3)
(4)
(5)
(6)
(7

yo
nu
re
it
in
de

e
p
p
e
c

pr

pr

ta

ap

for an item to be repeated using one of the following options:

(2) Repeat, no change.
(3) Repeat, louder.
(4) Repeat, slower.
(5) Repeat, louder and slower.
(6) Repeat, faster.
(7) Repeat, louder and faster.

To request a repetition simply state the option you feel will be most helpful to you either by number or by phrase. You may request up to three repetitions per item. If you do not require the item to be repeated or if you are finished with the instructions, request number one (no repeat or done).

There are 24 items in the set. At the end of each page (every six items) there will be a short pause before the item on the next page is presented. You will hear the speaker in your _____ ear. Along with the voice you will hear a noise composed of a group of people talking at the same time. I want you to ignore the noise completely. Pay attention only to the voice and carry out the instructions as accurately as you can.

Any questions? Let's try a practice test. The purpose of this set is to give you a chance to experience the task and to become familiar with the repetition options. In this set you will receive a sample of several different voices including the one you will be hearing. as with the regular test, you may request the repetition option you would like to have if given the chance. However, in the practice test the item will not be repeated. Simply carry out the instructions as accurately as you can.

The practice set was administered. This was the same practice tape used in Experiment 4. This process served to provide the subjects with exposure to both the experimental task and to the voice. Data acquisition time was approximately 45 minutes per subject.

Chapter III

Results and Discussion

Introduction

This study assessed a set of five commercially available digital speech synthesizers representing a range of technologies and costs. The goals of the study were to (1) develop instrumental and behavioral assessments that parallel selected aspects of a model of communication systems, (2) describe differences among speech synthesizers as revealed by various assessment methods and (3) compare rankings of speech synthesizers resulting from assessment methods. An additional goal was to define measures of "communication ergonomics" capable of reflecting the utility of speech synthesizers in tasks requiring exchange of information between machines and people.

Five experiments were devised to investigate the three major features of technical accuracy, semantic precision, and complex task performance. Technical accuracy was assessed through measurements of fundamental frequency (F_O) and format frequency (F_1, F_2, F_3) for each of nine vowels in a CVC context. Semantic precision was evaluated through a word-recognition-in noise task and a listening-comprehension-in noise task. Complex task performance was assessed using tests containing instructions of increasing durations and semantic complexity, with and without listener options for feedback control of speech sources.

un
ex
Ex
sc
ty
e
a

Fifty normal hearing persons were selected from a university population for participation in the four experiments requiring subjects (Experiments 2 to 4, N=12; Experiment, 5=14). Each subject received an audiological screening consisting of otoscopic examination, pure tone and tympanometric testing. Subjects were practiced in experimental tasks and speech sources prior to data acquisition.

The following experimental questions were asked:

(1) Do synthetic speech sources differ significantly from each other (and from human talkers) as a function of

a. acoustic structure of phonemes,
b. word-recognition measured in noise
c. comprehension measured in noise,
d. complex task performance?

(2) To what extent can b, c, and d (above) be predicted from a?

(3) What is the effect of listener-invoked communication repair upon the ability of listeners to accomplish complex tasks directed by speech synthesizers.

Technical Accuracy

Experiment 1: Acoustic Phonetic Analysis

Experiment 1 was designed to evaluate the technical accuracy of the five synthetic speech sources. Nine vowels in a CVC context were generated and recorded using the five synthesizers and a human control. The fundamental frequency and the first three formant frequencies were determined for each vowel using spectographic analysis. In order to calculate spectral loci coordinates it was necessary to have three format measurements for each vowel. Therefore, if a fo

ei

co

ar

we

vo

s

formant could not be identified on the spectrograph under either normal (black on white) or enhanced treatments (color coded for power) the vowel was rejected for use in further analysis. Three of the nine vowels recorded for analysis, were eliminated because of incomplete format data. Five vowels remained for [i, I, æ, ɛ, ɑ and ʌ] analysis.

Data Reduction

According to Miller (1984) a phoneme can be characterized on the basis of it's "spectral shape". Miller represents this shape as a single point in three dimensional space with the characteristics of $x = \log (F_3/F_2)$, $y = \log (F_2/F_0')$ and $z = \log (F_2/F_1)$. F_0' is defined as the fundamental frequency of the voice multiplied by a constant (1.5 greater for males). The first step in data reduction, therefore, was to determine the xyz coordinates for each vowel generated by the speech source. Tables 3.1 to 3.6 summarize the results of these calculations. Distance among the associated spectral loci were calculated using the formula:

$$D = [(x1 - x2)^2 = (y1 - y2)^2 = (z1 - z2)]^{.5}$$

Where, D distance is the square root of the sum of the squared differences along paired points indexed by three axes of the "Miller graph". The dependent variable of spectral loci distance were derived by summing distances across vowels. Calculations were initially conducted manually and were verified via Microsoft Excell, a spreadsheet program for use with Macintosh. Reliability between data sets was found

Tabl
calc
iden

Sour

Huma
DECt
Amig
Vot
Smo
Ech
Pet

ble 3.1. Speech spectra coordinates for the vowel [i]
lculated from algorithms suggested by Miller (1984) which
entify its placement in three dimensional space.

urce	x	y	z
man	.189	.205	.879
Ctalk	.162	.304	.689
iga	.185	.187	.750
trax PSS	.264	.525	.736
oothtalker	.152	.306	.910
ho II +	.150	.273	.477
terson & Barney	.188	.121	.928

ble 3.2. Speech spectra coordinates for the vowel [I]
lculated from algorithms suggested by Miller (1984) which
entify its placement in three dimensional space.

urce	x	y	z
Ctalk	.207	.391	.514
iga	.187	.371	.526
trax PSS	.264	.531	.736
oothtalker	.131	.422	.658
ho II +	.165	.449	.533
terson & Barney	.107	.284	.707

Table 3.3. Speech spectra coordinates for the vowel [ɛ] calculated from algorithms suggested by Miller (1984) which identify its placement in three dimensional space.

Source	x	y	z
Human	.171	.550	.368
DECtalk	.187	.501	.422
Amiga	.164	.512	.477
VOTRAX PSS	.136	.749	.433
Smoothtalker	.219	.422	.589
Echo II +	.164	.541	.415
Peterson & Barney	.129	.434	.540

Table 3.4. Speech spectra coordinates for the vowel [æ] calculated from algorithms suggested by Miller which identify its placement in three dimensional space.

Source	x	y	z
Human	.252	.520	.308
DECtalk	.199	.531	.380
Amiga	.204	.505	.368
VOTRAX PSS	.136	.749	.433
Smoothtalker	.244	.583	.385
Echo II +	.237	.439	.336
Peterson & Barney	.146	.539	.416

Table 3.5. Speech spectra coordinates for the vowel [ɑ] calculated from algorithms suggested by Miller (1984) which identify its placement in three dimemsional space.

Source	x	y	z
Human	.264	.579	.550
DECtalk	.315	.531	.286
Amiga	.352	.539	.243
VOTRAX PSS	.284	.760	.268
Smoothtalker	.109	.747	.528
Echo II +	.121	.541	.636
Peterson & Barney	.350	.539	.174

Table 3.6. Speech spectra coordinates for the vowel [ʌ] calculated from algorithms suggested by Miller (1984) which identify its placement in three dimemsional space.

Source	x	y	z
Human	.325	.406	.373
DECtalk	.339	.501	.285
Amiga	.270	.432	.348
VOTRAX PSS	.284	.711	.301
Smoothtalker	.355	.547	.335
Echo II +	.273	.541	.329
Peterson & Barney	.302	.516	.269

to be on the order of ± .001. Table 3.7 summarizes these distances and the absolute sums of the distances. Fundamental frequency and formant center frequency data extracted from the spectograms of human and synthetic speech were compared to archival data for male talkers (N=33) taken from Peterson and Barney (1952) and plotted in the same manner. These findings were later used in correlational analyses to determine the relationship between data obtained in Experiment 1 and in Experiments 2,3 and 4.

Statistical Methods and Tools For Experiments 2 Through 5

Subject performance for Experiments 2, 3, 4 and 5 was indexed, in part, by percent-correct scores. Other measure included completion time in seconds (Experiments 2 and 3) and percent of communication repair options selected (Experiment 5). Means standard, deviations and ranges were calculated for all indices. Because the percent-correct variables were based upon different numbers of items (e.g. 50 for SPIN full-list; 25 for SPIN subtests), and because percent-correct indices often are skewed, these data were transformed prior to further analysis. Studebaker (1985) suggested the use of a "rationalized" arcsin transform (R) to overcome the problems associated with percentage data. Advantages cited for using R include (1) the correction of correlation between means and variances typical of percentages, (2) linearization of data relative to the variance, thus allowing for direct comparisons between all parts of the performance range, (3) the provision of percentage-like numbers for

Table 3.7.Distance value for vowels in three dimensional space from Peterson & Barney (1952) vowel measurements as calculated using xyz coordinates derived from Miller (1984) scaled x 100.

Source	Vowel						
	[i]	[ɪ]	[ɛ]	[æ]	[ɑ]	[ʌ]	SUM
Human	12.02	12.22	21.26	15.21	51.25	15.31	127.27
DECtalk	30.44	24.20	14.80	6.46	13.29	4.26	93.45
Amiga	20.16	21.60	10.64	8.23	8.80	11.96	81.38
PSS	47.02	29.35	33.27	21.06	20.28	19.92	170.90
Smooth	18.85	14.84	10.30	9.43	44.89	9.06	107.90
Echo II +	47.72	24.66	16.85	15.66	51.87	7.13	163.90
SUM	176.20	126.87	107.13	76.05	190.39	67.64	---

purposes of discussion, and (4) permitting variance and standard deviation to be interpreted in the same manner as percentage data.

Data transform was implemented via Microsoft Excell. Percent-correct scores were converted to proportions by dividing by 100. The proportions (p) were then entered into the first column of the spreadsheet. Proportions were converted to radians by the formula:

$$T = \arcsin \sqrt{(p * N/N+1)} + \arcsin \sqrt{(p * N + 1/N+1)}.$$

Where T is the arcsin transform expressed in radians and N is the number of items tested. The rationalized arcsin transform (RAST) was calculated in the next column as recommended by Studebacker (1985) using the formula:

$$R=46.47324337T - 23$$

Where R is a linear transformation of T expressed in Raus. Rau is defined as a quasi-physical unit with no physical dimension. The resulting transformed scores expressed in raus were used in subsequent statistical analysis. It is important to note that values may exceed 100 or be less than 0.

Analysis was accomplished using two software programs. Descriptive statistics (means, standard deviations and ranges) and correlations were generated using StatView (Abacus Concepts, 1986). ANOVAs', Newman-Keuls' post hoc test of mean differences and simple effects were calculated using CLR ANOVA (Clear Lake Research, 1985). Criteria for significance for all statistical tests was $p \leq 0.05$.

Semantic Precision

Experiment 2: Word-recognition

The SPIN test was used to evaluate semantic accuracy via word-recognition in sentences. Subjects were asked to listen to a set of fifty sentences presented in the presence of a twelve-voice babble noise (+8 S/B) and to repeat the last word of the sentence. Speech sources were presented in randomized orders and SPIN lists were counterbalanced within sources. Error counts were made separately for the full-list and half-list high and low-predictability key words and the results compared for accuracy of count. Error counts were always conducted at least twice regardless of full and half-list agreement. In addition, random selection of seven score sheets (10%) indicated no difference in error count between first and later scorings.

Description. Tables 3.8 and 3.9 summarize mean percent-correct, standard deviations and ranges of performance as a function of speech source for full-list SPIN test, high-predictability and low-predictability word scores. Tables 3.10 and 3.11 display mean arcsin transformed percent-correct scores, standard deviations and ranges. Overall, scores for speech synthesis sources were poorer than the human control. Aside from natural speech the highest mean score occurred with DECtalk (81.6) and the lowest with Echo II + (28.8). There was considerable variation among synthesized speech sources as reflected by the ranges.

Table 3.8. Mean percent-correct scores, standard deviations, and ranges for SPIN full-list test results as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	90.6	2.4	88-99
DECtalk	81.6	7.0	70-92
Amiga	66	7.6	54-76
Votras PSS	51.1	8.2	36-56
Smoothtalker	36.5	7.8	26-54
Echo II+	28.8	7.2	16-40

Table 3.9. Mean percent-correct scores, standard deviations, and ranges for SPIN high-predicatibility and low-predictability test results as a function of speech source (N=12).

Word Predictability		Mean	S.D.	Range
Human				
	High-predictability	100	0	100 -100
	Low-predictability	81.3	4.9	76 - 88
DECtalk				
	High-predictability	94.6	4.9	88 -100
	Low-predictability	70.3	11.1	48 - 84
Amiga				
	High-predictability	81	9.3	64 -92
	Low-predictability	50.6	9.8	36 -68
Votras PSS				
	High-predictability	62.6	12.4	32 -76
	Low-predictability	39.6	13.3	20 -60
Smoothtalker				
	High-predictability	45.8	12.7	24 -68
	Low-predictability	26	8.6	12 -40
Echo II+				
	High-predictability	30	12	12 -52
	Low-predictability	17.8	9.3	2 -32

Table 3.10. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for SPIN full-list test results as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	93.1	3.8	89.0 -98.0
DECtalk	81.7	8.5	68.0 -95.0
Amiga	65.0	7.3	53.6 -74.8
Votras PSS	51.0	7.5	37.0 -64.8
Smoothtalker	37.4	7.3	27.2 -53.6
Echo II+	29.8	7.3	16.0 -40.8

Table 3.11. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for SPIN high-predicatibility and low-predictability test results as a function of speech source (N=12).

Word Predictability		Mean	S.D.	Range
Human				
	High-predictability	113.83	0	113.0-113
	Low-predictability	80.3	5.6	74.3-88.1
DECtalk				
	High-predictability	113.2	11.3	88 -113
	Low-predictability	69.1	10.9	48.2 -83.1
Amiga				
	High-predictability	80.3	10.7	62.2 -93.8
	Low-predictability	50.6	8.9	37.3 -66.4
Votras PSS				
	High-predictability	61.6	11.5	33.5 -74.3
	Low-predictability	40.3	12.5	21.3 -59.0
Smoothtalker				
	High-predictability	46.1	11.6	25.6 -66.4
	Low-predictability	27.2	8.9	11.8 -41.0
Echo II+				
	High-predictability	31.0	11.7	11.8 -51.7
	Low-predictability	17.7	11.7	-5.2 -33.5

Figures 3.1 and 3.2 are histograms illustrating the full-list and high-predictability and low-predictability percent correct scores. Each bar represents mean percent-correct or mean arcsin transformed percent-correct scores for twelve subjects. Standard deviations are provided in the lower histograms. Figures 3.3 and 3.4 display the functions for the transformed scores. The figures demonstrate the rank ordering of the speech sources as a function of word-recognition scores and the differences between the high and low-predictability scores for each speech source. The mean differences across speech sources between high and low-predictability scores were 21.4 % and 27 raus for the percent correct and arcsin transformed percent-correct scores respectively.

Statistical procedures. To investigate the differences noted in SPIN scores across speech sources a one-way, repeated measures analysis of variance (ANOVA) was performed. Table 3.12 summarizes the results of ANOVA. Results were significant. Thus, word recognition varied as a function of speech source. To establish the extent of the differences among speech sources a Newman-Keuls' test of paired comparisons was performed revealing that each mean differed significantly from every other mean.

A two-way, mixed-effects ANOVA was used to evaluate the main effects of speech source and SPIN subtests, as well as the interaction between the two effects. Both main effects were repeated, i.e. each subject received both SPIN subtests

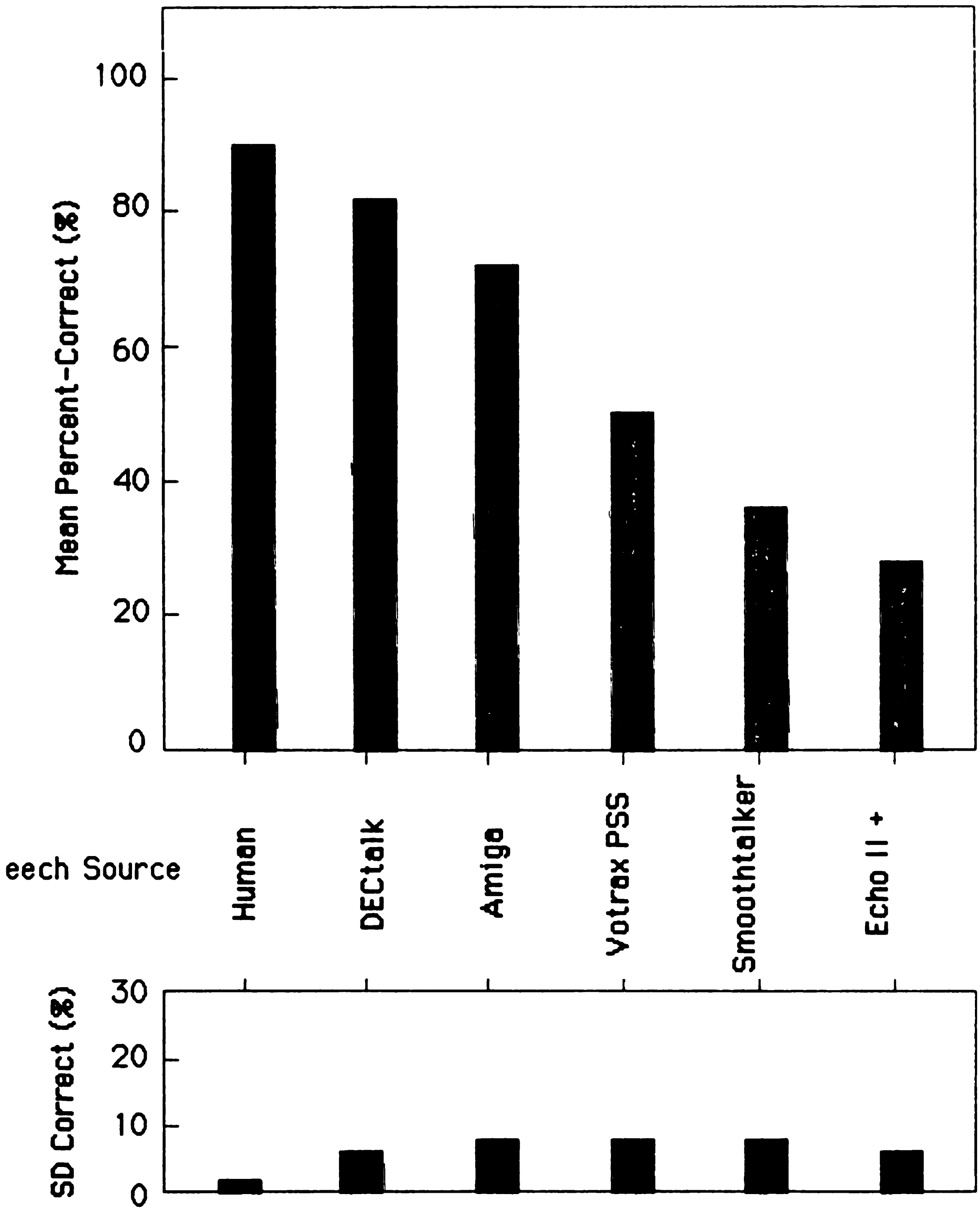


Figure 3.1. Mean percent-correct scores and standard deviations for SPIN full-list test results as a function of speech source. Each bar represents observations of 12 normal-hearing adult subjects tested monaurally. The lower histogram denotes ±1 standard deviation.

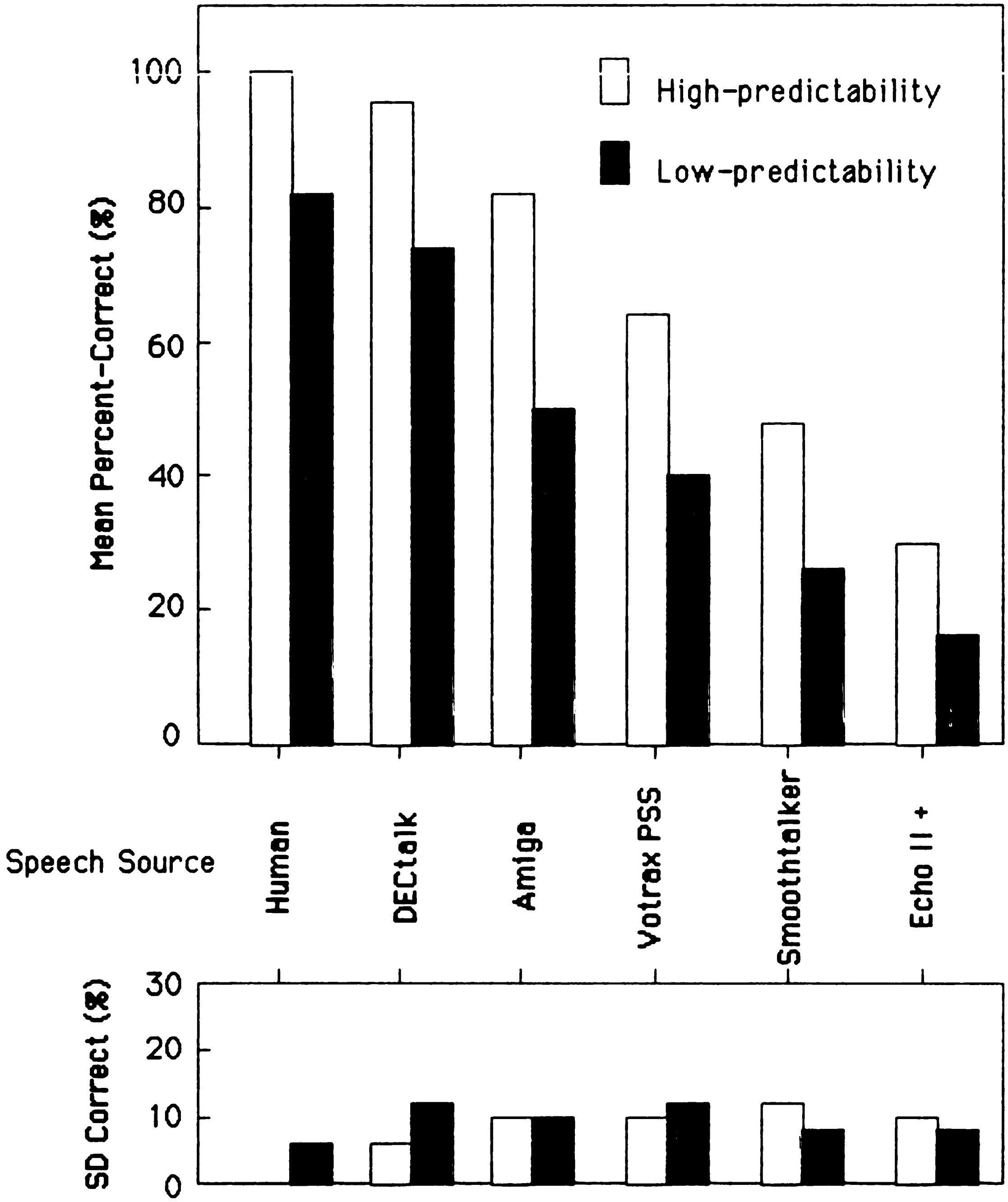


Figure 3.2. Mean percent-correct scores and standard deviations for SPIN high-predictability and low-predictability word sets as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally. The lower histogram denotes ± 1 standard devitation.

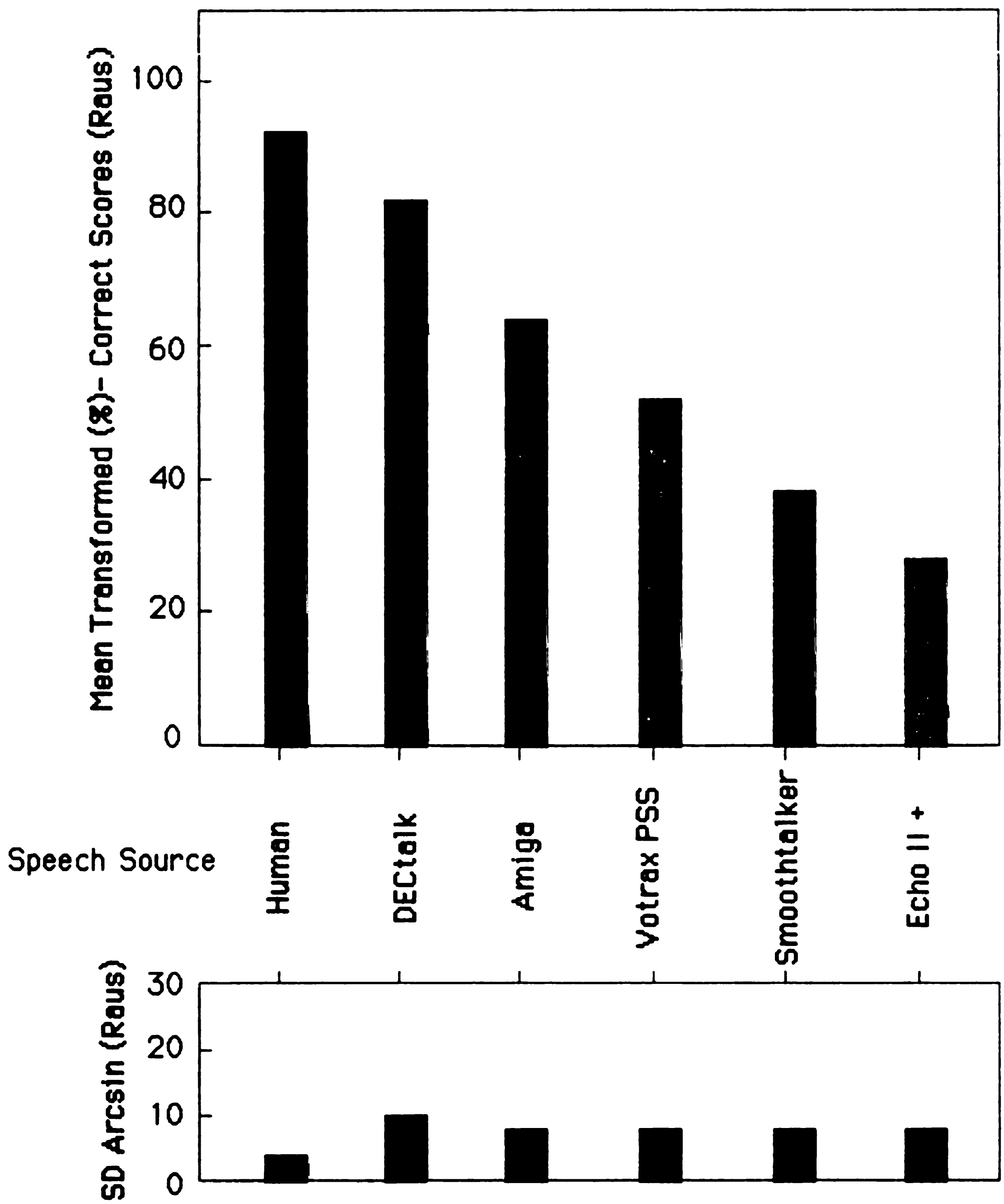


Figure 3.3. Mean arcsin transformed percent-correct scores and standard deviations for SPIN full-list test results as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally. The lower histogram denotes ± 1standard deviation.

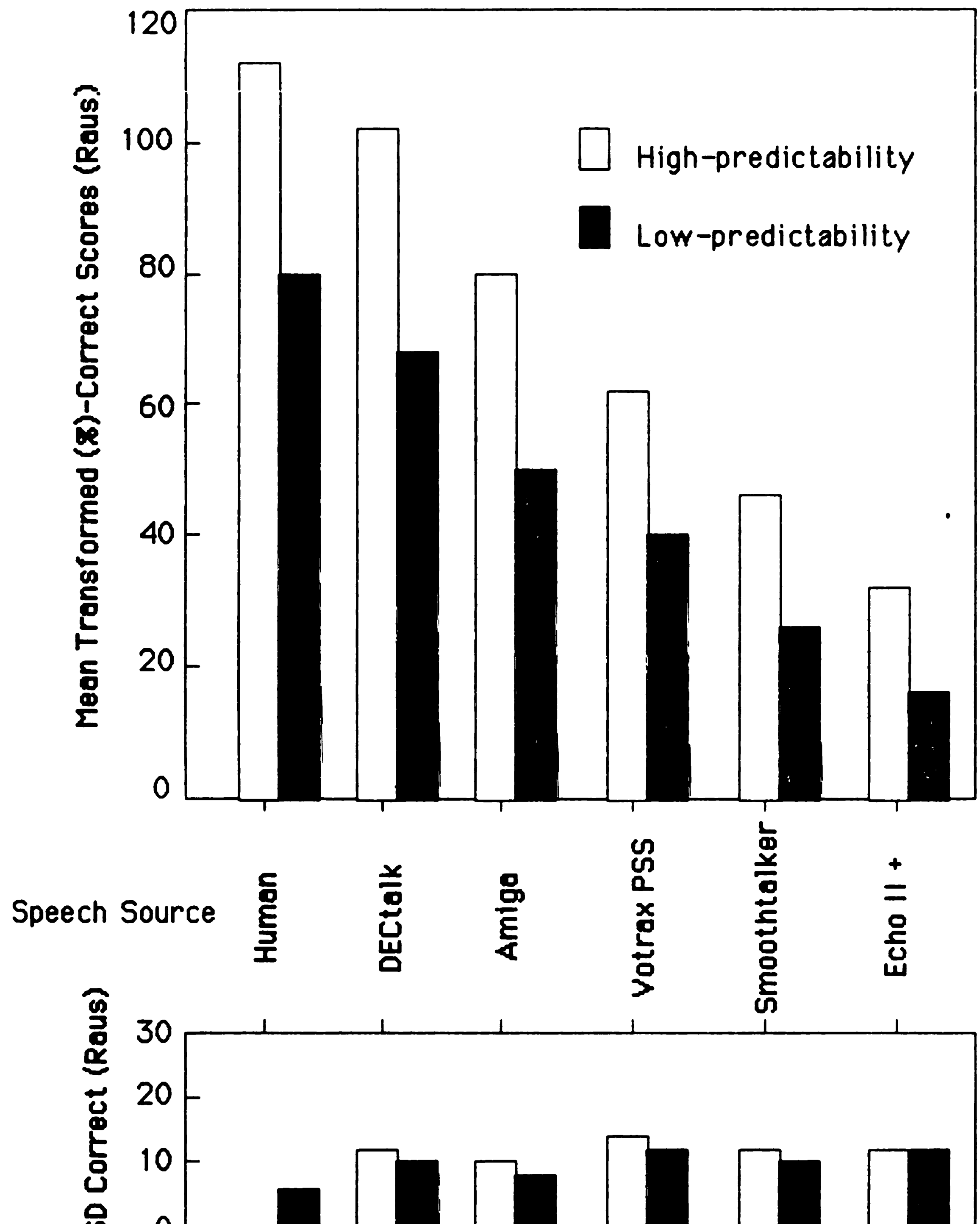


Figure 3.4. Mean arcsin transformed percent-correct scores and standard deviations for SPIN high-predictability and low-predictability word sets as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally. The lower histogram denotes ±1 standard deviation.

Table 3.12 One-way within-subject analysis of variance of full-list SPIN test results.

Source	dF	SS	MS	F	p
Subjects	11	475.944	41.6		
Speech Source	5	36516.944	7303.389	143.885	<.001
Error	55	2791.722	50.759		

(low-predictability and high-predictability key words) as produced by each of the six speech sources (DECtalk, Amiga, VOTRAX PSS, Smoothtalker, Echo II+, and a male human control). Table 3.13 displays the results of these calculations. The main effects of speech source and linguistic predictability (SPIN high and low-predictability half-lists) proved to be significant, as was the interaction of source and predictability. Thus, speech sources resulted in different word recognition performance with regard to linguistic predictability and high and low-predictability scores differed with regard to speech source.

To further assess the interaction between the two main effects the Newman-Keuls' test of pairwise comparisons was conducted. Figure 3.5 illustrates these results by using a heavy solid line to identify those variables whose mean pairs did not differ significantly. Five mean pairs did not demonstrate significant differences including; Human (81.3 R) and Amiga (81 R) high-predictability, DECtalk (70.3 R) low and VOTRAX PSS (62.6 R) high-predictability, Amiga (50.6 R) low and Smoothtalker (45.8 R) high-predictability, VOTRAX PSS low (39.6 R) and Smoothtalker (45.8 R) high-predictability, and Smoothtalker (28 R) low and Echo II+ (30 R) high-predictability. Table 3.14 lists the results of simple effects tests on the interaction of the two main effects. All F-ratios were significant.

Implications. The experimental question associated with Experiment 2 was whether synthetic speech sources produced

Table 3.13. Two-way within-subject analysis of variance of SPIN test results with Main Effects of speech source and word predictability.

Source	dF	SS	MS	F	p
Subjects	11	1256.971	114.27		
Speech Source	5	94310.696	18862.139	144.374	<.001
Error	55	7185.611	130.647		
Predictability	1	22196.779	22196.779	479.619	<.001
Error	11	509.08	46.28		
Source/ Predictability	5	1997.125	399.452	4.57	.0015
Error	55	4806.611	87.393		

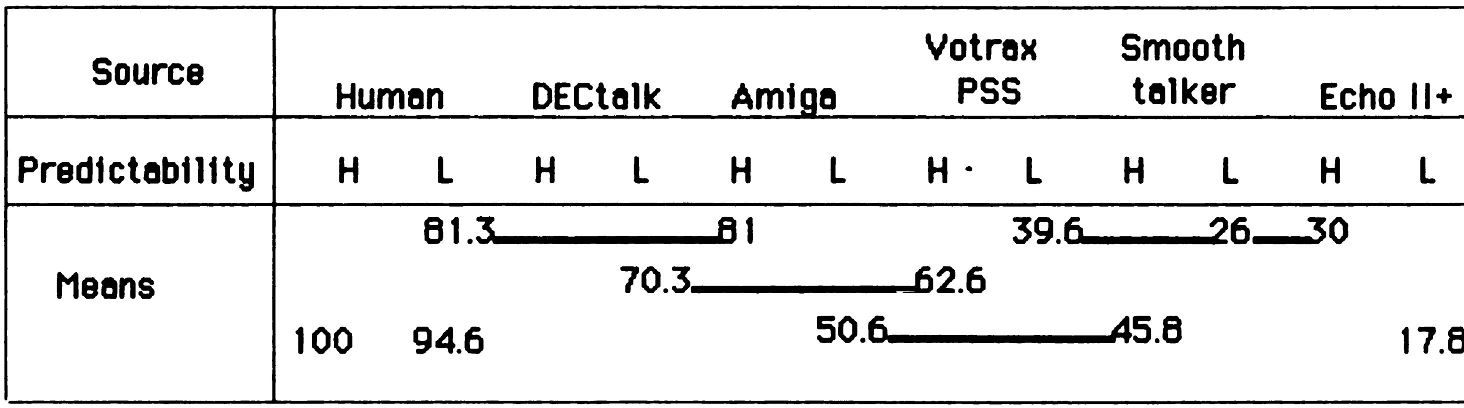


Figure 3.5. Illustration of Newman-Keuls' test of pairwise comparisons of SPIN test transformed percent-correct scores for high-predictability and low-predictability words as a function of speech source. Nonsignificant mean pairs are connected by solid lines.

Table 3.14. Simple effects for the variables of speech source and linguistic preditability based on transformed percent-correct scores of the SPIN of high-and low-predictability key word subtests.

Effect	MSn	DFn	DFe	MSe	F	p
Source at high-predictability	12294.83	5	55	113.631	108.199	<.000
Source at low-predictability	6966.729	5	55	104.406	66.7625	<.000
Predictability at Human	6728.141	1	11	15.972	421.237	<.000
Predictability at DECtalk	6157.447	1	11	65.058	94.646	<.000
Predictability at Amiga	5370.341	1	11	67.014	80.138	<.000
Preditability at VOTRAX PSS	2725.336	1	11	186.096	19.645	<.000
Predictability at Smoothtalker	2151.963	1	11	109.707	19/616	<.000
Predictability at Echo II+	1060.675	1	11	39.398	26.922	<.000

differences in word recognition in noise. On the basis of the present outcomes it can be said that significant differences do exist between all speech sources for word recognition in sentences. These findings are consistent with those obtained in previous research (Pisoni; Mirenda & Beukelman, 1987; Hoover et al. 1987), though it should be noted not all of the devices employed here have been used by other investigators. Only DECtalk, VOTRAX and Echo II+ have been used in other word recognition studies.

High and low-predictability subtests of the SPIN made it possible to assess the interaction between speech source and linguistic predictability. Once again the results confirm those obtained in similar studies, at least with regard to DECtalk (Greene and Pisoni, 1988). Performance improved increased message redundancy as reflected by the higher scores for high-predictability key words.

Pairwise comparisons provided some interesting patterns of mean differences. For example, the Amiga high-predictability subtests did not differ significantly from the Human low-predictability scores and were significantly better than the DECtalk low-predictability scores. This is in contrast to full-list results in which Amiga proved to be significantly poorer than both the human and DECtalk. This suggests that a device with mid-range intelligibility can be used successfully in high-predictability communication situations, while even a high quality device will exhibit markedly poorer performance in comparison to a human source

in the absence of adequate message redundancy. It is likely similar outcomes would occur for sets of test materials that differ in word familiarity or word frequency.

Experiment 3: Listening Comprehension

Factual passages were used to evaluate semantic accuracy via a measure of listening comprehension. The passages were presented in the presence of a twelve-voice babble noise (+ 8 S/B) and were immediately followed by a multiple choice test regarding passage content. Speech sources were presented in randomized orders and passages were counter balanced within sources. The time required to complete each multiple choice test was recorded. Error counts for the multiple choice tests were repeated a minimum of two times. A random selection of seven tests (10%) indicated no differences in error count.

Description. Table 3.15 contains mean percent-correct, mean standard deviations and ranges as a function of speech source for multiple-choice comprehension test scores. Corresponding transformed percent-correct scores (Raus), standard deviations and ranges are provided in Table 3.16. The speech synthesizer demonstrating the highest and lowest mean tramsformed scores were DECtalk (82.3 R) and Echo II + (49 R). Examination of the ranges reveals a wide variation in performance within all speech sources. Table 3.17 shows mean multiple choice test completion time, standard deviations and ranges as a function of speech source. Test completion times demonstrated little variation as a function

Table 3.15. Mean percent-correct scores, standard deviations, and ranges for multiple-choice comprehension test results as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	82	16.4	50-100
DECtalk	82.3	8.9	66.6-100
Amiga	74.6	21.3	33.5-100
Votras PSS	72.5	16.6	50-100
Smoothtalker	55.9	13.6	33.3-75
Echo II+	49	15	25 -75

Table 3.16. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for multiple choice comprehension test results as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	81.9	19.1	10 -107.2
DECtalk	80.0	11.0	64 -107.2
Amiga	74.3	22.2	36 -107.2
Votras PSS	71.2	17.5	50 -107.2
Smoothtalker	55.1	11.6	35 -71.58
Echo II+	49.1	12.7	28 -71.5

Table 3.17. Mean completion time, standard deviations, and ranges of multiple choice comprehension tests (in seconds) as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	38.2	11.9	20.2 -56.6
DECtalk	45.5	8.8	26.1 -59.3
Amiga	45.7	15.2	18.0 -80.9
Votras PSS	43.4	14.6	21.3 -73.6
Smoothtalker	46.3	9.0	32.2 -60.0
Echo II+	46.5	12.6	31.0 -74.1

of synthesized speech source (43.4 to 46.5 secs). Table 3.18 displays arcsin transformed percent-correct test scores across all speech sources as a function of test version. Figures 3.6, 3.7 and 3.8 are histograms displaying mean test scores, mean transformed percent-correct scores and mean multiple choice test completion time. In can be seen that the rank ordering of speech sources on the basis of transformed percent-correct comprehensive test scores remains the same as in Experiment 2 (Human, DECtalk, Amiga, VOTRAX, Smoothtalker and Echo II+).

Statistical Procedures. Passages used as experimental stimuli were originally designed for use with human speakers. The question arises regarding the equivalency of the passages produced by speech synthesizer. Though the versions were counterbalanced across synthesizers, nonequivalent versions could effect statistical outcomes. Noting the presence of higher trans formed scores for tests three (77.4 R) and four (79.7 R), a one-way within-subject repeated measures ANOVA was performed. The F-ratio failed to reach significance, thus suggesting the equivalency of the multiple choice comprehension tests across speech sources. It is assumed,therefore, that results were not confounded by differences among versions of the measurement device.

Differences amongtransformed multiple choice test scores were analyzed using a one-way repeated measures ANOVA. Results were significant. Indicating that listening comprehension varies as a function of speech source. Table

Table 3.18. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for the multiple choice comprehension tests (N=12).

Multiple Choice Test	Mean	S.D.	Range
Test 1	66.9	10.9	54.1 -85.9
Test 2	63.4	22.0	39.5 -107.2
Test 3	77.4	21.4	28.4 -107.2
Test 4	79.7	24.1	35.8 -107.2
Test 5	65.6	13.3	50.0 -84.3
Test 6	58.4	19.4	35.8 -85.9

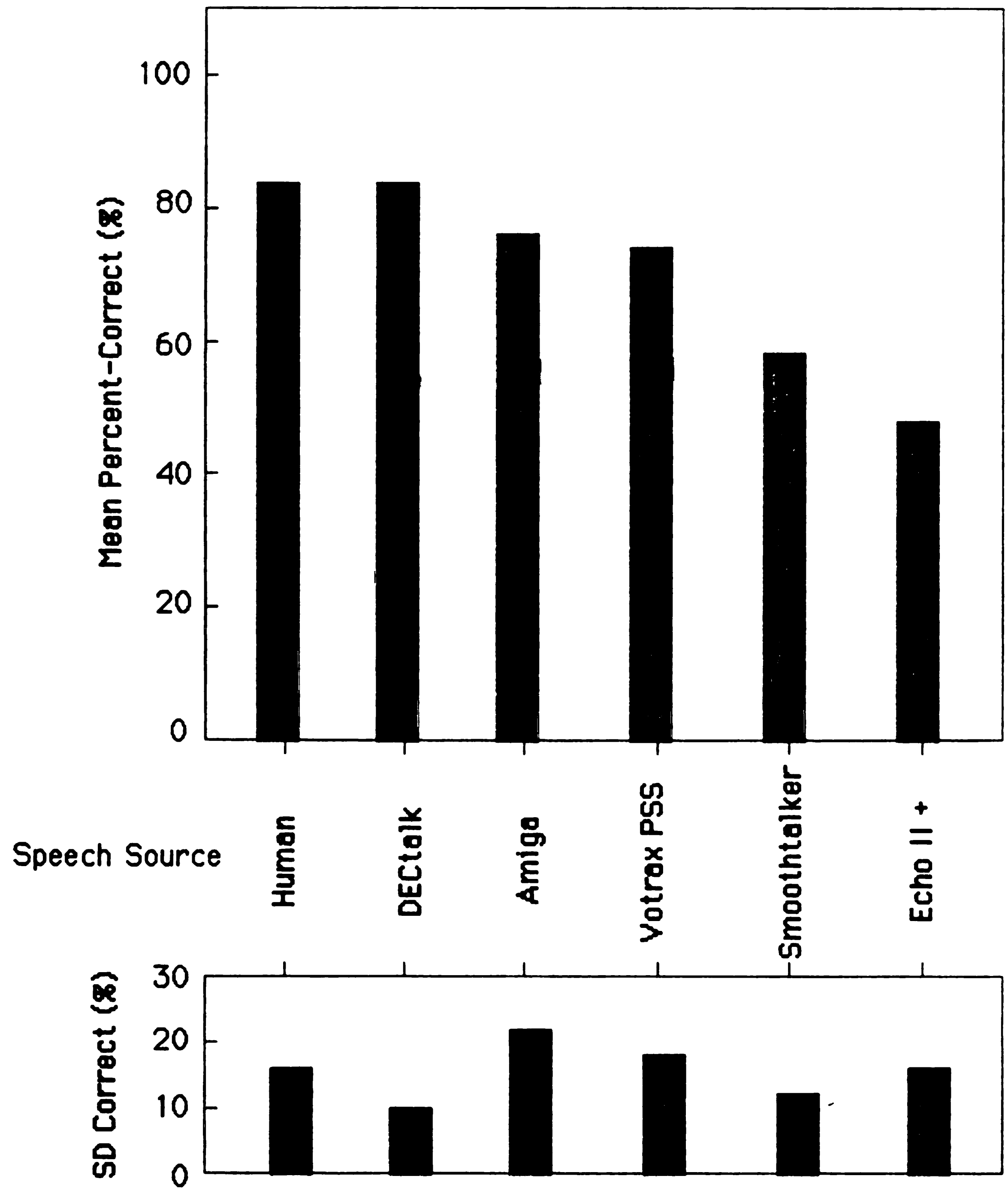


Figure 3.6. Mean percent-correct scores and standard deviations for multiple choice comprehension tests as a function of speech source. Each bar represents observations of 12 normal-hearing subjects. The lower histrogram denotes ±1 standard deviation.

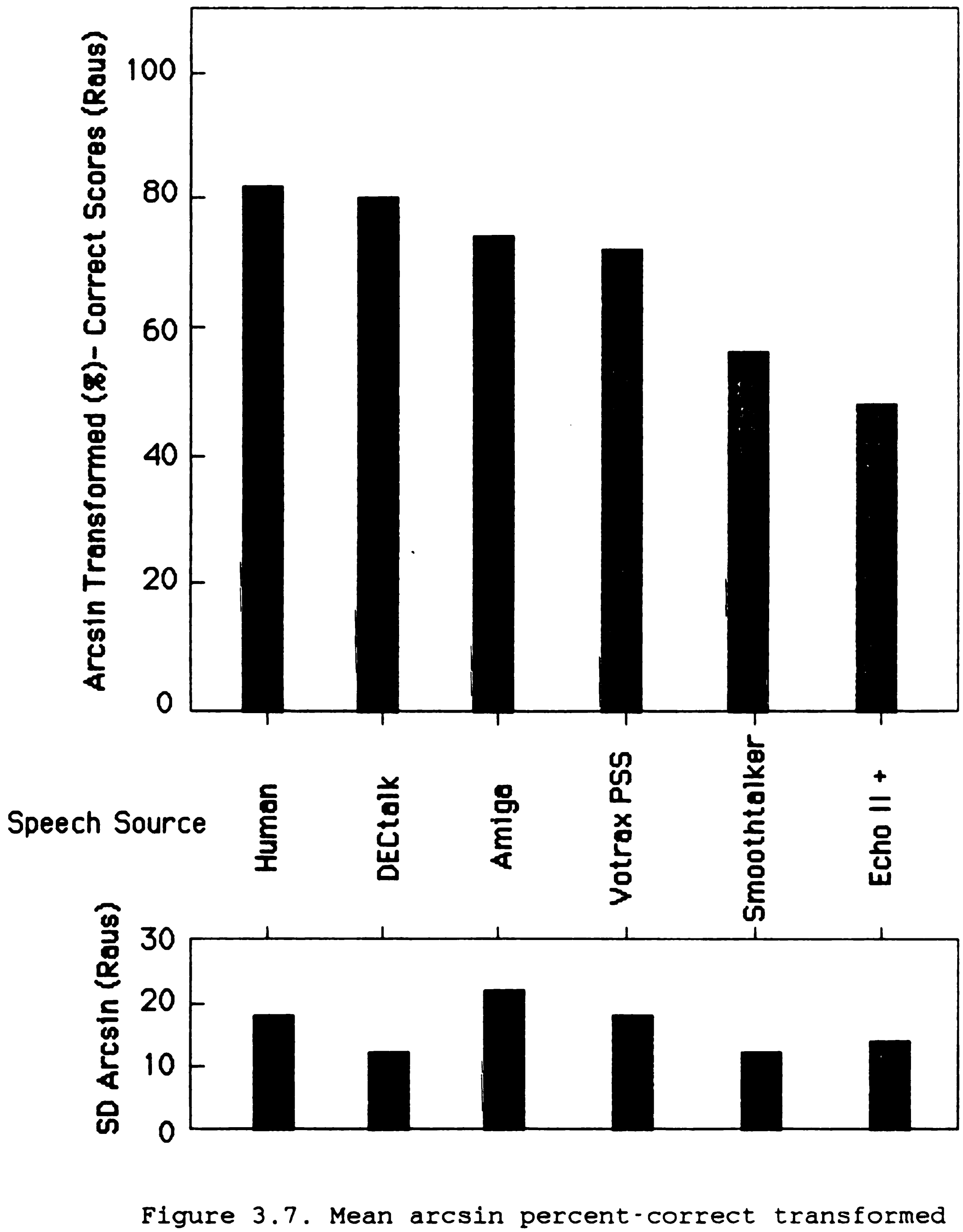


Figure 3.7. Mean arcsin percent-correct transformed scores and standard deviations for multiple choice comprehension tests as a function of speech source. Each bar represents observations of 12 subjects. The lower historam denotes ±1 standard deviation.

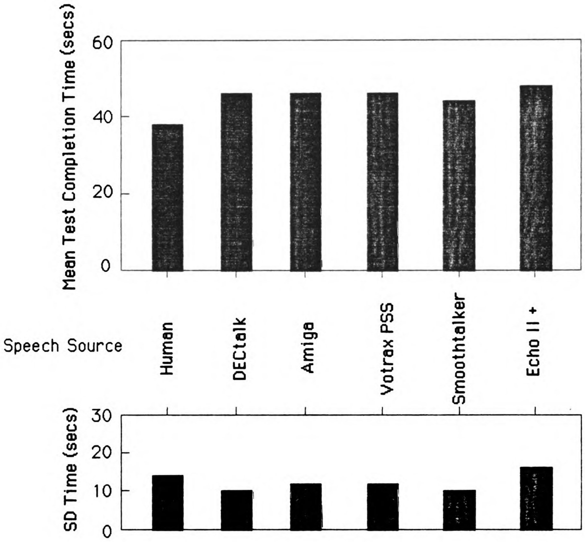


Figure 3.8. Mean multiple choice test completion time and standard deviations in seconds as a function of speech source. Each bar represents observations of 12 normal-hearing subjects. The lower histogram denotes ±1 standard deviation.

3.19 summarizes these results. A post hoc analysis revealed six pairs of means that did not differ. These included; Human/DECtalk, Human/Amiga, Human/VOTRAX PSS, DECtalk /Amiga, DECtalk/ VOTRAX PSS, and Amiga/VOTRAX PSS. Figure 3.9 illustrates the results of the Newman-Keuls' test of paired comparisons. It can be seen that Smoothtalker and ECHO II + were the only speech sources which varied significantly from all other speech sources: listening comprehension was significantly poorer for these speech synthesizers.

A one-way repeated measures ANOVA also was performed to evaluate differences in mean multiple choice test completion times. No significant differences between speech sources were noted. Thus, comprehension test completion time did not effectively differentiate speech sources.

Time alone can be considered an index of the ergonomics of human-computer interaction, however, the relationship between performance and time can also be viewed as an ergonomic metric. Ideally, efficient information exchange promotes speed, but not at the cost of accuracy. In the present study efficiency was calculated by dividing mean accuracy (percent-correct comprehension) by mean time-to-test completion (seconds). Figure 3.10 illustrates outcomes for this derived variable. Ranking according to percent-correct per second follows the order of (1) Human (2.14), (2) DECtalk (1.75), (3) VOTRAX PSS (1.64), (4) Amiga (1.62), Smoothtalker (1.19) and Echo II + (1.05). This ordering is very similar to that for comprehension test scores. However,

Table 3.19. One-way within-subject analysis of variance of multiple choice arcsin transformed percent-correct scores for speech source.

Source	dF	SS	MS	F	p
Subjects	11	1727.745	157.068		
Speech Source	5	10871.231	2174.246	7.577	<.000
Error	55	15781.647	286.939		

Source	Human	DECtalk	Amiga	Votrax PSS	Smooth talker	Echo II+
Means	81.9 ——	80.0	74.3 ——	71.2	55.1	49.1
	81.9 ——	——	74.3			
	81.9 ——	——	——	71.2		
			74.3 ——	71.2		
		80.0 ——	——	71.2		

Figure 3.9. Illustration of Newman-Kuels' test of pairwise comparisons for transformed multiple choice test comprehension scores as a funtion of speech source. Nonsignificant mean pairs are connected by solid lines.

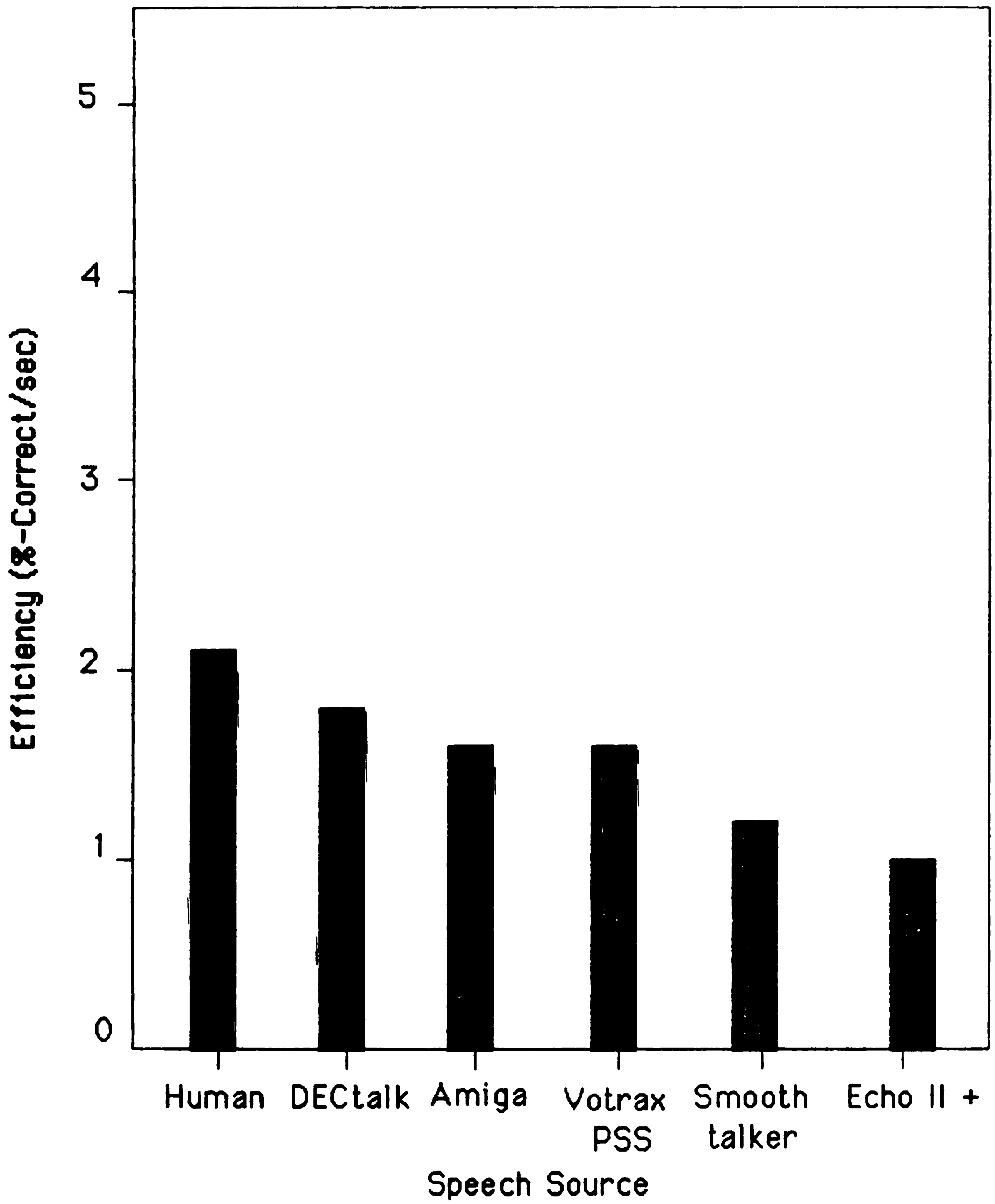


Figure 3.10. Efficiency of six speech sources for a listening comprehension task. Efficiency was calculated by dividing mean percent-corect scores for a multiple-choice comprehension test by the time (in seconds) required to complete the test.

differences between the Human talker and DECtalk become more apparent as task completion time is taken into account. Amiga and VOTRAX exchanged positions as a result of the PSS's slightly smaller mean test completion time. Even so, the efficiency ratio equalizes these two sources, thus emphasizing the similarity in their performance for this task.

Implications. The experimental question asked in with Experiment 3 was whether listening comprehension in noise differed as a function of speech source. Significant differences do exist between some synthetic speech sources, even though the effects of source on comprehension were less systematic than the effects of source on word recognition. These results are in contrast to the findings of McHugh (1976), Schwab, Nusbaum & Pisoni 1(985) and Greene and Pisoni (1988). None of these studies demonstrated significant differences between synthesized speech sources or between synthesized speech sources and a human control. One possible explanation might lie in the use of different stimuli. However, the passages employed in Experiment 3 were originally designed for use with grade school children. Those used by Greene and Pisoni (1988) were taken from adult reading comprehension tests. Logically, the greater difficulty associated with an adult task should be more effective in differentiating between speech sources. This was not case. Another possibility lies in the choice of speech synthesizers. In previous research the MITalk -79 has been shown (Greene, Logan and Pisoni, 1986) to have a high degree of intelligibility, thus a listening comprehension task might not be sufficient to demonstrate small differences in performance. This is supported by the present findings reported in that differences were noted only for the poor quality synthesizers in relation to each other and to the other speech sources. Mid-range and high intelligibility devices did not demonstrate significant differences. However, McHugh (1976) used an early (and it is assumed, poorer) version of VOTRAX and was unable to demonstrate differences between the speech synthesis device and a human control. Finally, the use of the twelve-voice-babble in the present study very likely increased the processing difficulty sufficiently to reveal those devices most affected by reduced redundancy. As noted previously, studies such as Chial (1973) and Pisoni and Koen (1981) have demonstrated the negative effect of noise on the perception of synthesized speech.

In summary, both measures of semantic accuracy demonstrated differences as a function of speech source. Rank ordering of speech sources based on percent-correct scores resulted in the same hierarchy for both experiments. However, word recognition in sentences as measured by the Revised SPIN Test resulted in a pattern of significant differences among all speech sources. This suggests the R-SPIN is a more sensitive measure of semantic precision than the combination of factual passages and multiple choice comprehension tests.

Task Performance

Experiment 4: Oral Instructions (Without Repair Options)

Task performance was evaluated using the Multiple Instructions Test. Subjects were asked to follow a set of instructions of varying complexity presented monaurally in conjunction with a twelve-voice babble noise (+10 S/B). Six alternate forms of the MIT were generated by all speech sources. Order of speech synthesis was randomized across subjects and versions of the MIT were counterbalanced across speech sources. Times-to-completion of each item was recorded and summed to generate total test completion time.

Subjects responded by marking clear plastic sheets overlayed on graphic respons forms. Because it was necessary to erase the subject responses after each session, copies were made of each response sheet. However, all scoring was done from the original response forms and verified during the duplicaiton process. Error counts were then made from the score sheet. Counts were verified by a volunteer during the process of tabulating errors.

Description. Tables 3.20 to 3.24 give summary statistics for percent-correct scores, andarcsin transformed percent-correct scores (raus), of each level of the Multiple Instructions Test as a function of speech source. Scores ranged from 71.3% (DECtalk) to 43.2% (Echo II +) for the synthetic sources as compared to 77.2 % for the human

Table 3.20. Mean percent-correct, standard deviations, and ranges for total scores of the Multiple Instructions Test as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	77.2	6.9	66.6 -88
DECtalk	71.3	9.1	52.3 -83
Amiga	67.9	10.3	42.8 -80.9
Votras PSS	67.2	7.1	55.9 -79
Smoothtalker	57.2	5.0	48.8 -66.6
Echo II+	43.2	14.0	14 -61.9

Table 3.21. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for total scores of the Multiple Instructions Test as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	75.9	7.4	65.0 -88
DECtalk	69.9	8.9	52.0 -81.9
Amiga	66.6	9.7	43.5 -79.5
Votras PSS	65.8	6.8	55.2 -77.4
Smoothtalker	56.5	4.5	48.9 -65.0
Echo II+	43.5	13.2	14 .4-60.7

Table 3.22. Mean percent correct, standard deviations, and ranges for all complexity levels of the Multiple Instructions Test as a function of speech source (N=12).

	MIT Level	Mean	S.D.	Range
Human				
	Level A	97.9	3.7	91.6 - 100
	Level B	87.4	17.2	44.4 -100
	Level C	73.1	10.4	58.3 -87.7
	Level D	68	8.1	56.5 -80
DECtalk				
	Level A	95.1	4.3	91.6 -100
	Level B	85.8	12.9	61.0 -100
	Level C	67.0	14.5	45.3 -91
	Level D	57.3	11.9	33.3 -70
Amiga				
	Level A	92.9	10.0	66.6 -100
	Level B	72.9	18.5	33.3 -100
	Level C	60.6	11.9	37.5 -75
	Level D	62.4	11.9	43.3 -67.6
Votras PSS				
	Level A	85.2	16.3	50.0 -100
	Level B	77.3	13.4	55.0 -94
	Level C	60.5	13.3	45.0 -83
	Level D	58.9	11.9	43.0 -86.6
Smoothtalker				
	Level A	76.9	10.7	58.0-91.6
	Level B	60.4	12.9	33.3 -83
	Level C	46.6	14.2	29.1 -70.8
	Level D	56.4	11.6	33.3 -70
Echo II+				
	Level A	59.6	24.3	16.0 -91.6
	Level B	50.3	19.4	22.2 -83
	Level C	38.7	16.5	1.2 -62.5
	Level D	38.1	16.8	10.0 -66.6

Table 3.23. Mean transformed percent-correct scores, standard deviations, and ranges for all complexity levels of the Multiple Instructions Test as a function of speech source (N=12).

	MIT Level	Mean	S.D.	Range
Human				
	Level A	100.7	7.6	87.9 -104.9
	Level B	87.7	19.4	45.4 -104.9
	Level C	69.6	9.5	56.7 -83.3
	Level D	64.9	6.9	55.3-75.5
DECtalk				
	Level A	95.0	8.7	97.9 -104.9
	Level B	70.9	18.1	36.3 -104.9
	Level C	64.5	12.7	46.2 -87.2
	Level D	56.0	9.7	36.3 -66.4
Amiga				
	Level A	93.5	13.6	63.5 -104.9
	Level B	73.8	12.4	54.0 -91.1
	Level C	58.7	9.8	39.8 -70.8
	Level D	60.3	9.9	44.5 -72.3
Votras PSS				
	Level A	84.1	17.8	50.0 -104.9
	Level B	73.8	12.4	54.0 -91.1
	Level C	58.7	11.2	45.9 -78.4
	Level D	58.4	11.5	43.7 -86.3
Smoothtalker				
	Level A	73.2	10.3	56.4 -87.9
	Level B	58.7	10.9	36.3 -78.4
	Level C	47.2	11.6	32.7 -67.1
	Level D	55.2	9.4	36.3 -66.4
Echo II+				
	Level A	58.5	21.2	21.1 -87.98
	Level B	50.4	16.3	26.5 -78.4
	Level C	39.9	15.6	.41-60.1
	Level D	39.8	14.7	13.9 -63.5

control. Scores for all speech sources declined as MIT complexity level increased. Table 3.24 provides the arcsin transformed percent-correct scores for the experimental versions of the MIT. Tables 3.25 and 3.26 contain mean test completion time, standard deviations and ranges for the total MIT and for each level of the test.

Figure 3.11 to 3.14 are histograms illustrating mean test scores, transformed scores, and standard deviations for the total test and for each level. These illustrate both the rank ordering of the synthesizers resulting from the MIT scores and the differences between level of item complexity. Different levels of the MIT are identified by bar coding within each speech source. Figure 3.15 displays the arcsin transformed percent-correct scores for the experimental versions of the MIT across versions of the test. Figures 3.16 and 3.17 summarize the results of item completion for the total test and for complexity level.

Statistical Procedures.

A one-way repeated measures ANOVA was performed to determine if differences existed among versions of the MIT for all speech sources. The F-ratios were not significant indicating that the six versions of the MIT did not differ.

Table 3.27 contains the findings of a two-way, mixed effects ANOVA performed on transformed percent-correct scores to evaluate the significance of the main effects of speech source and MIT complexity level, as well as the interaction between these two main effects. Differences

Table 3.24. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for experimental versions of the Multiple Instructions Test across all speech sources (N=12).

MIT Test	Mean	S.D.	Range
Test 2	64.8	10.4	42.8 -79.6
Test 3	65.9	12.1	39.1 -99.0
Test 4	60.9	12.4	29.7 -74.3
Test 5	61.6	15.9	33.5 -81.9
Test 6	60.1	17.4	14.4 -79.6
Test 8	69.5	13.3	45.5 -85.5

Table 3.25. Mean completion time, standard deviations, and ranges for the total Multiple Instructions Test (in seconds) as a function of speech source (N=12).

Speech Source	Mean	S.D.	Range
Human	104.0	33.2	104 -150.3
DECtalk	116.5	17.4	88.4 -145.1
Amiga	120.0	28.6	87.0 -166.8
Votras PSS	118.9	26.4	73.2 -170.5
Smoothtalker	118.2	23.1	91.0 -166.3
Echo II+	106.8	30.6	52.7 -165.9

Table 3.26. Mean item completion time in seconds, standard deviations, and ranges in for all levels of the Multiple Instructions Test as a function of speech source (N=12).

	MIT Level	Mean	S.D.	Range
Human				
	Level A	2.7	.96	1.1 -4.3
	Level B	4.2	1.1	2.5 -5.9
	Level C	6.1	1.1	4.0 -8.0
	Level D	6.1	1.6	3.4 -9.0
DECtalk				
	Level A	2.9	.65	2.1 -4.3
	Level B	4.2	.86	2.7 -5.5
	Level C	6.0	1.4	4.4 -7.4
	Level D	6.0	1.2	4.1 -8.0
Amiga				
	Level A	3.3	.68	2.3 -4.6
	Level B	4.4	1.1	3.0 -6.3
	Level C	5.5	1.4	3.7 -7.9
	Level D	6.6	2.0	3.8 -10.5
Votras PSS				
	Level A	3.2	.59	2.2 -4.2
	Level B	4.5	1.0	3.1 -6.5
	Level C	5.8	1.7	4.0 -9.1
	Level D	6.4	1.1	4.8 -8.7
Smoothtalker				
	Level A	3.3	.8	1.6 -4.8
	Level B	4.6	.92	3.8 -6.6
	Level C	5.8	1.3	3.6 -7.9
	Level D	5.9	1.5	4.1 -8.6
Echo II+				
	Level A	3.5	.86	2.3 -5.6
	Level B	4.0	1.0	2.3 -6.2
	Level C	4.9	1.8	2.4 -8.8
	Level D	5.3	1.9	1.7 -8.2

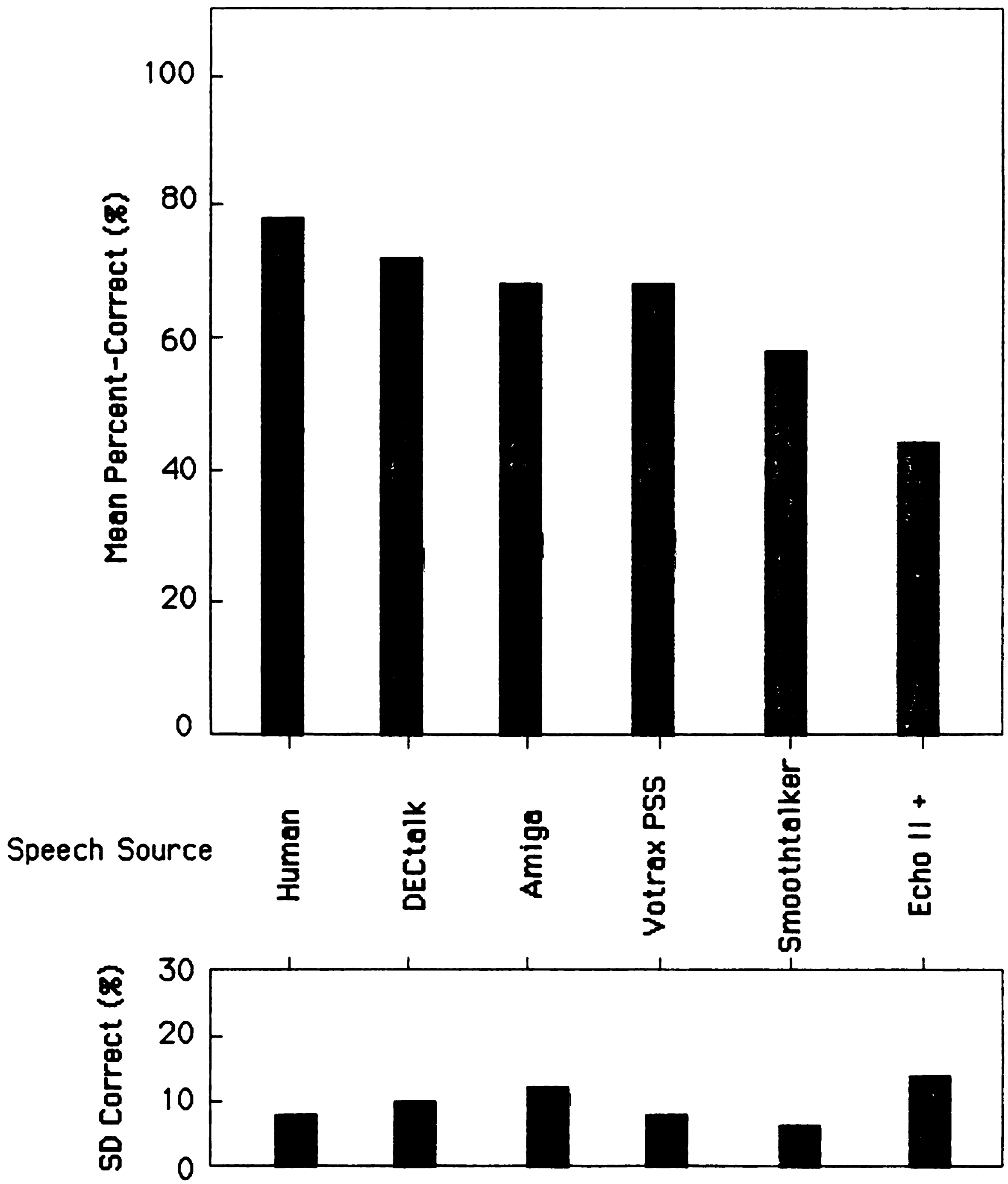


Figure 3.11. Mean percent-correct scores and standard deviations for Multiple Instructions Test as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

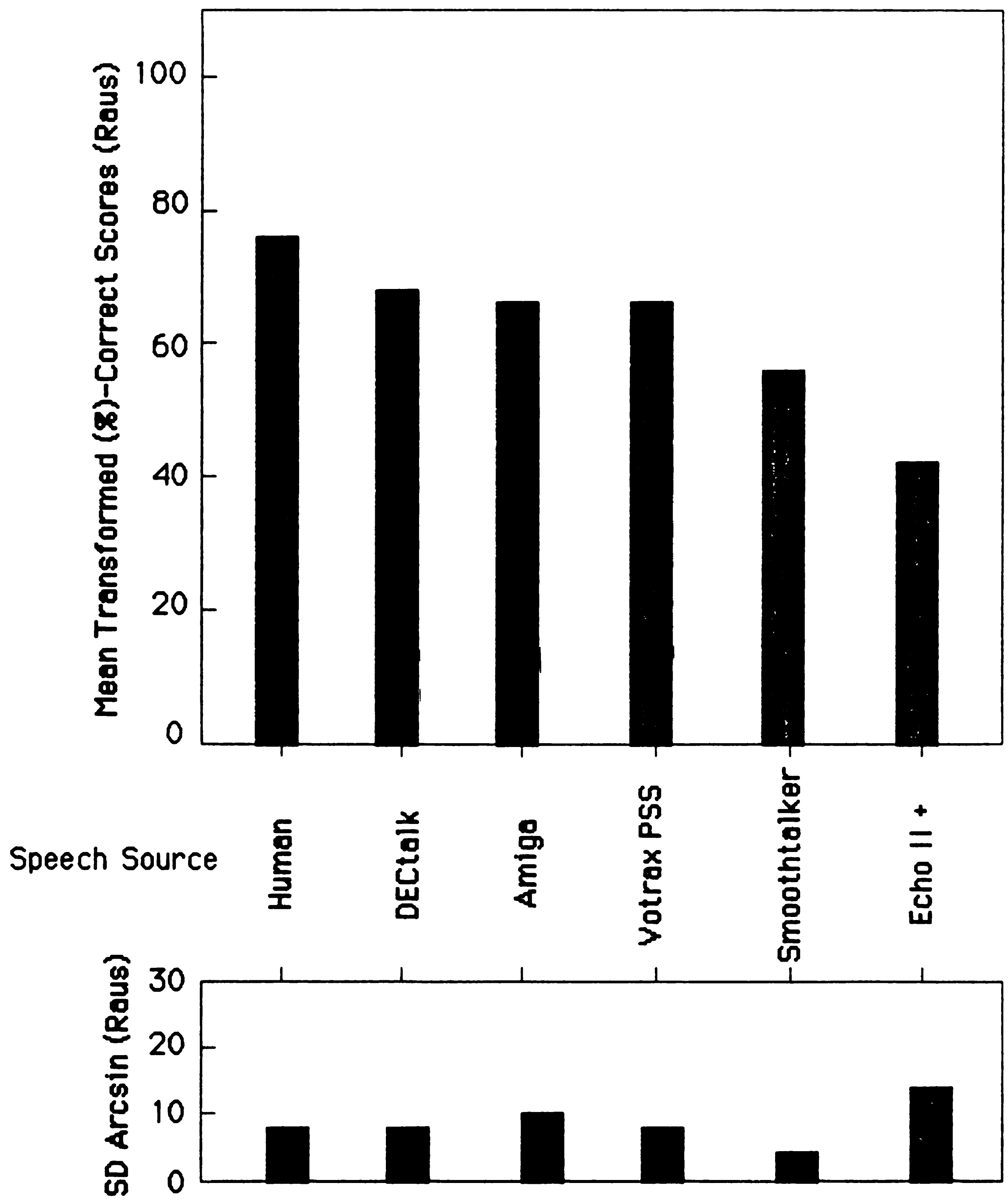


Figure 3.12. Mean transformed percent-correct scores and standard deviations for Multiple Instructions Test as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

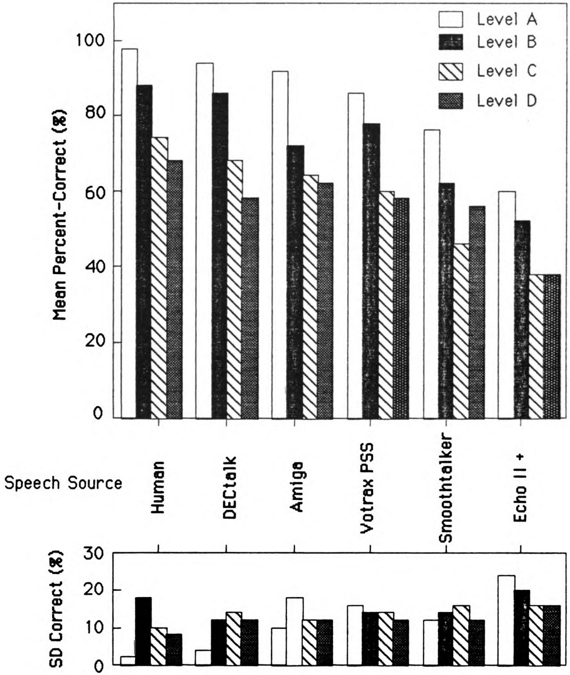


Figure 3.13. Mean percent-correct scores and standard deviations for all levels of the Multiple Instructions Test as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

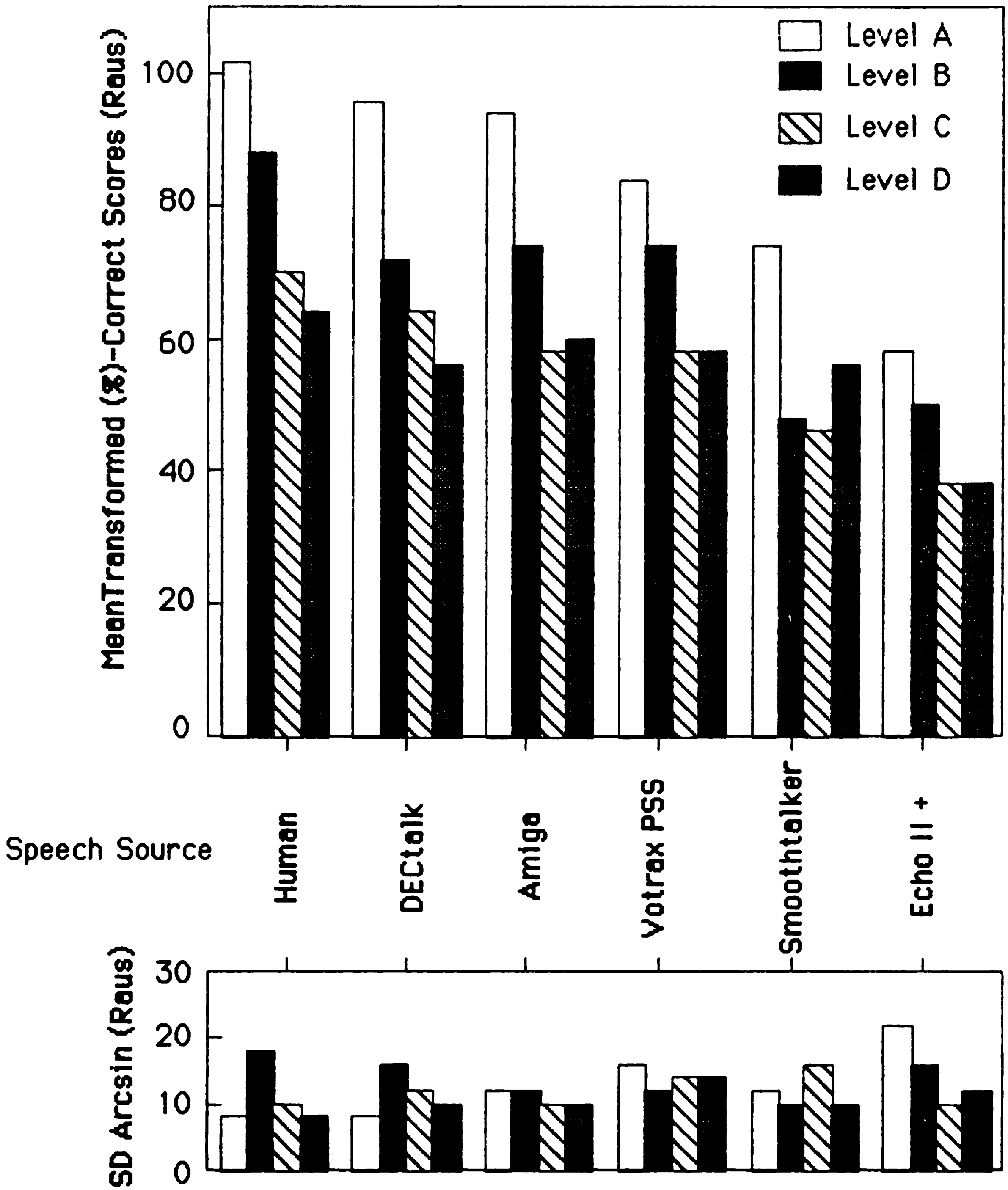


Figure 3.14. Mean transformed percent-correct scores and standard deviations for all levels of the Multiple Instructions Test as a function of speech source. Each bar represents 12 normal-hearing subjects tested monaurally.

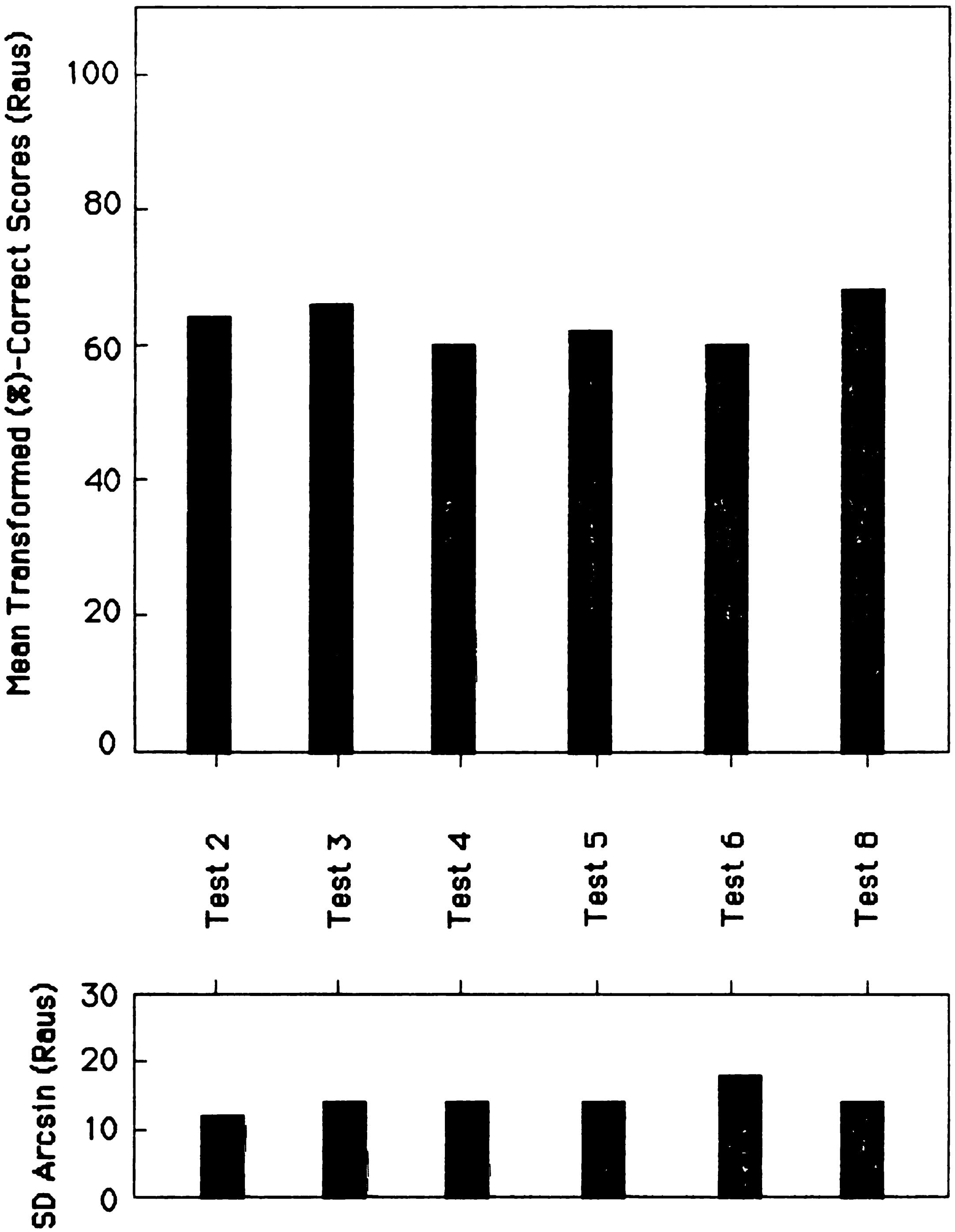


Figure 3.15. Mean transformed percent-correct scores and standard deviations for the six experimental versions of the Multiple Instructions Test. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

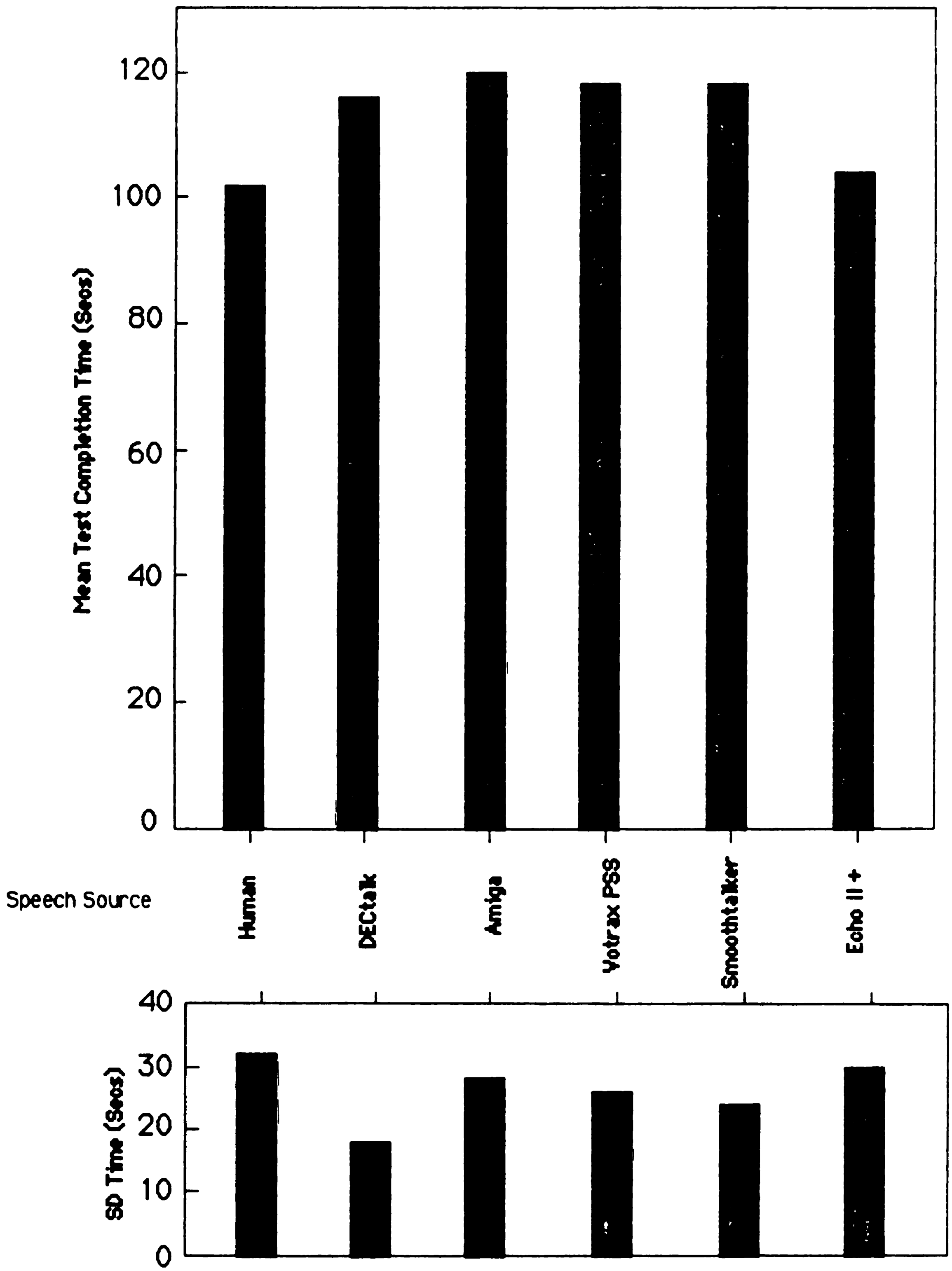


Figure 3.16. Mean completion time in seconds and standard deviations all items of the Multiple Instructions Test as a function of speech source. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

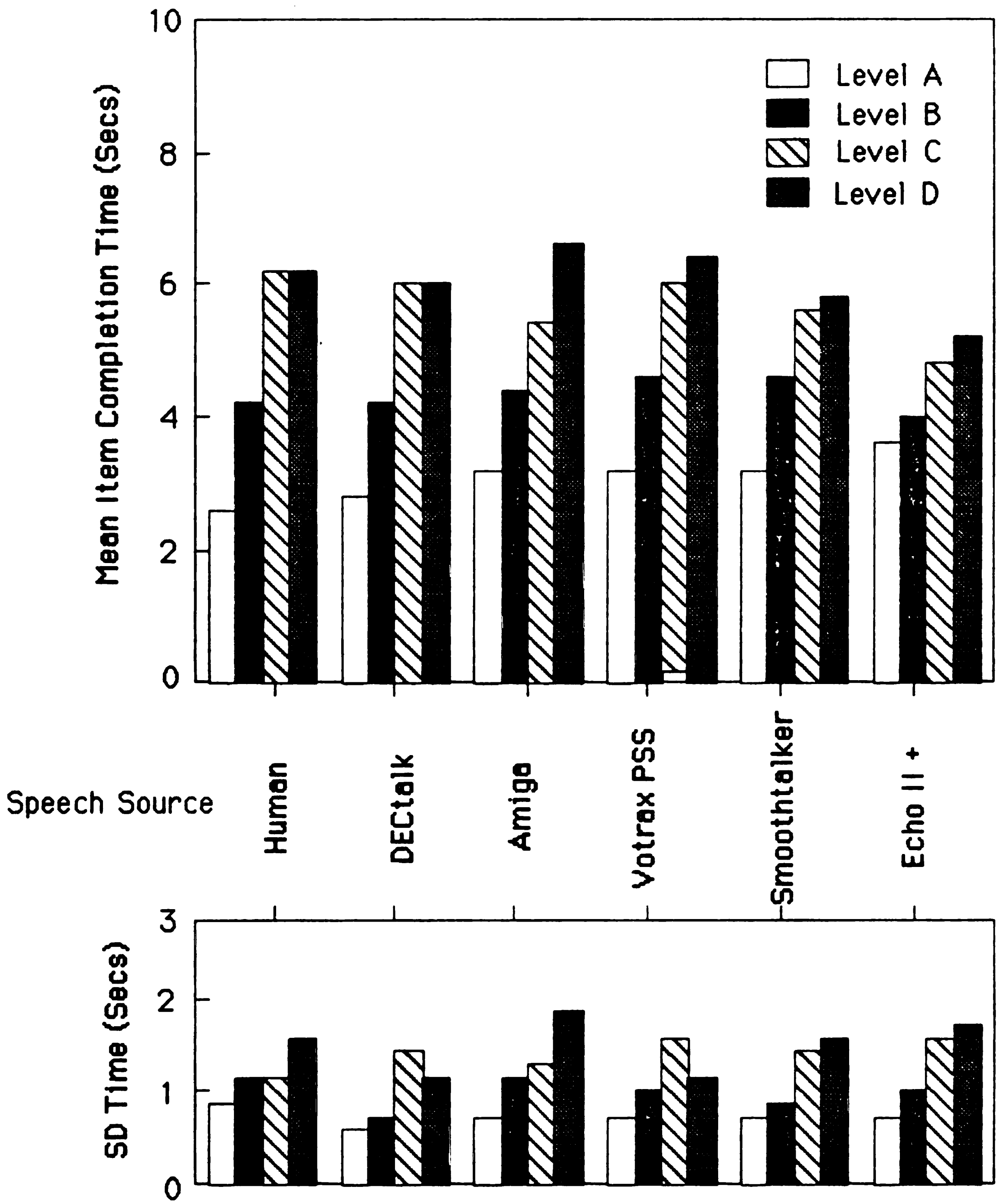


Figure 3.17. Mean item completion time in seconds and standard deviations for each level of the Multiple Instructions Test as a function of speech source. Each bar represents 12 normal-hearing subjects tested monaurally.

Table 3.27. Two-way within-subject ANOVA of Multiple Instructions Test transformed percent-correct scores.

Source	dF	SS	MS	F	p
Subjects	11	7217.445	655.677		
Speech Source	5	33511.346	6702.269	30.730	<.000
Error	55	11995.455	218.099		
Level	3	38587.081	12862.360	131.990	<.000
Error	33	3215.844	97.450		
Source/Level	5	3559.759	237.317	1.692	.0568
Error	165	23141.124	140.249		

between means for the two main effects of speech source and complexity level were significant. The interaction between source and level failed to meet the criterion for significance. Thus, MIT performance varied as a function of speech source and of item complexity level, but the two effects cannot be said to interact.

To further analyze the interaction of means (raus) Newman-Keuls' test of paired comparisons were performed on the main effects of speech source and MIT level of complexity. Figures 3.18 and 3.19 illustrate these results. Three pairs of means did not demonstrate significant differences. These included DECtalk/Amiga, DECtalk/VOTRAX PSS and Amiga/VOTRAX PSS. Thus, the findings suggest subject task performance did not differ significantly for DECtalk, Amiga, and VOTRAX PSS. However, task performance scores did differ among these synthesizers and the human source, as well as among these and Smoothtalker and Echo II+. Levels C and D of the MIT did not differ significantly, but Levels A, B, and C. di differ from each other.

Table 3.28 lists the results of simple effects comparisons between speech source and level and between level and speech source. All F-ratios were significant suggesting that each speech source produced significant differences at each level and each level produced significant differences for each source.

To assess the differences between means (raus) of speech source and item completion time and the interaction between

Table 3.28. Simple effects for the variables of speech source and level of item complexity based on transformed percent-correct scores of the MIT..

Effect	MSn	DFn	DFe	MSe	F	p
Source at Level A	3019.966	5	55	211.899	14.252	<.000
Source at Level B	2049.838	5	55	208.252	9.843	<.000
Source at Level C	1462.540	5	55	188.352	12.358	<.000
Source at Level D	881.876	5	55	100.344	8.789	<.000
Level at Human	3281.689	3	33	123.609	26.549	<.000
Level at DECtalk	3370.514	3	33	147.848	22.797	<.000
Level at Amiga	3115.125	3	33	112.357	27.725	<.000
Level at VOTRAX PSS	1875.982	3	33	163.696	11.460	<.000
Level at Smoothtalker	1422.726	3	33	120.129	11.843	<.000
Level at Echo II+	982.911	3	33	131.058	7.500	<.000

Source	Human	DECtalk	Amiga	Votrax PSS	Smooth talker	Echo II+
Means	80.7	71.6 ——	71.6		56.5	43.5
			71.6 ——	65.8		
		71.6 ——	——	65.8		

Figure 3.18. Illustration of Newman-Kuels' test of pairwise comparisons for Multiple Instructions Test arcsin transformed percent-correct scores as a function of speech source. Nonsignificant mean pairs are connected by solid lines.

Complexity Level	Level A	Level B	Level C	Level D
Means	**84.2**	**69.2**	**56.4** ——	**55.7**

Figure 3.19. Illustration of Newman-Keuls' Test of pairwise comparisons of mean arcsin transformed percent-correct scores for the Multiple Instructions Test for the main effect of item complexity level. Nonsignificant mean pairs are connected by solid line.

them, a two-way within-subject repeated measures ANOVA was performed with the main effects of speech source and MIT mean item completion time at each level of complexity. Table 3.29 summarizes the results. The F-ratio for the effect of speech source was not significant. Thus, item completion time did not differentiate between speech sources. The F-ratio for the main effect of item completion time at level of complexity was significant, as was the interaction between speech source and item completion time. The Newman-Keul's test revealed the only Level C and D did not differ according to item completion time. These results are provided in Figure 3.20.

Measures of simple effects are shown in Table 3.30. Only one of the paired means for speech source and item completion time was significant. However, item completion time differed for each level within speech source. Thus, mean item completion time varied significantly for MIT level of complexity for all speech sources.

Efficiency (percent-correct per second) was calculated for the MIT using total test scores and total test completion time. Results are displayed in Figure 2.21. Once again, the speech sources can be ranked according to the data; (1) Human (.74), (2) DECtalk (.612), (3) Amiga (.574), (4) VOTRAX PSS (.565), (5) Smoothtalker (.483) and (6) (Echo II + .404). As in Experiment 3, the inclusion of time clarifys distinctions between human and synthesized speech sources. Thus, task completion times can be useful even in the absence of

Table 3.29. Two-way within-subject analysis of variance of Multiple Instructions Test arcsin transformed percent-correct scores with Main Effects of speech source and time per level.

Source	dF	SS	MS	F	p
Subjects	11	190.788	17.344		
Speech Source	5	8.941	1.788	.981	.4375
Error	55	100.224	1.822		
Time	3	406.127	135.376	65.639	<.001
Error	33	68.060	2.062		
Source/Time	15	26.071	1.738	3.651	<.001
Error	165	78.544	.476		

Complexity Level	Level A	Level B	Level C	Level D
Means	**3.1**	**4.36**	**5.73 ——**	**—6.1**

Figure 3.20. Illustration of Newman-Keuls' Test of pairwise comparisons of mean item completion time in seconds for the Multiple Instructions Test for the effect of item complexity level. Nonsignificant mean pairs are connected by solid line.

Table 3.30. Simple effects for the variables of speech source and level of complexity for item completion time (in seconds)of the MIT..

Effect	MSn	DFn	DFe	MSe	F	p
Source at Level A	1.534	5	55	.461	3.328	.011
Source at Level B	.687	5	55	.647	1.062	.329
Source at Level C	2.276	5	55	1.021	2.248	.062
Source at Level D	2.507	5	55	1.131	2.218	.065
Time at Human	32.193	3	33	.755	41.546	<.000
Time at DECtalk	30.193	3	33	.775	42.336	<.000
Time at Amiga	32.409	3	33	.713	44.123	<.000
Time at VOTRAX PSS	23.787	3	33	.808	29.434	<.000
Time at Smoothtalker	17.736	3	33	.727	24.382	<.000
Time at Echo II+	7.747	3	33	.698	11.319	.001

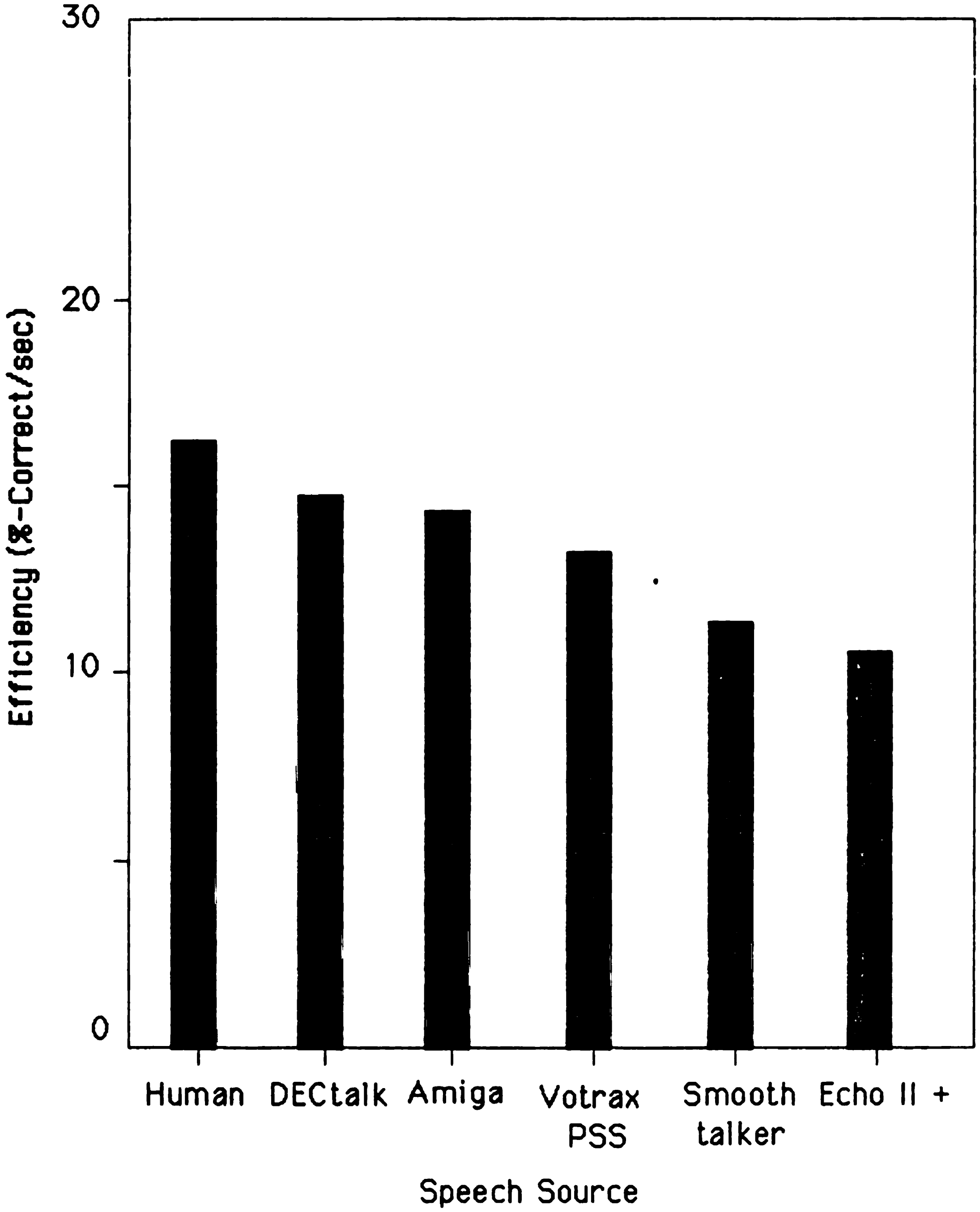


Figure 3.21. Efficiency of six speech sources for the Multiple Instructions Test. Efficiency was calculated using percent-correct total test scores and completion times in seconds. Each bar represents observations of 12 normal-hearing subjects.

significant differences between sources. Using this information in combination with performance data is useful in rating the efficiency with which tasks are performed under the direction of various speech sources.

Implications. Experiment 4 asked whether task performance differed significantly as a function of speech source. It can be said that significant differences do exist between performance as a function of some speech sources. On the basis of task scores speech sources could be grouped in a manner suggested by Greene,Logan and Pisoni (1986) into (1) natural speech (human control) , (2) high to mid-range speech (DECtalk, Amiga, VOTRAX PSS), and (3) low quality speech (Smoothtalker, Echo II+). Once again, overall order is the same as that seen in both Experiments 2 and 3. Thus, the speech sources employed in this study demonstrated consistency with regard to relative intelligibility when assessed at different points in the communication model. However, the differences in pattern of grouping reinforces the need to consider the task when selecting devices for specific applications.

Significant differences between item complexity levels for both transformed scores and time indicate test design was successful in creating a hierarchy of difficulty through the first three levels. Level D did not vary significantly from level C in either score or completion time. Thus, the present array of level D items did not prove to be sufficiently difficult to delineate a fourth level of

performance. These items could be removed from the test or redesigned to discover which components might increase the difficulty beyond that already seen in level C.

Experiment 5: Oral Directions (With Repair Option)

Subjects were asked to follow a set of instructions of varying complexity presented monaurally in conjunction with a twelve-voice babble noise (+10 S/B) using test 4 of the MIT. Test items were generated using the VOTRAX PSS text-to-speech system. Following presentation of a test item, subjects were allowed to select from seven repair options; (1) no repeat, (2) repeat no change, (3) repeat louder, (4) repeat slower, (5) repeat louder and slower, (6) repeat faster, and (7) repeat louder and faster. A maximum of three repair options were allowed per item and the total test. The number of each type and the total for the task as a whole were generated via the program used to present stimulus items. These values were used to verify experimenter counts at the end of each session.

As with Experiment 4, copies were made of each response sheet, scoring was done from the original forms and verified during the copy process. Error counts were then made from the score sheet.

Three sets of percentages were calculated from the repair options recorded; (1) the percent of each repair option selected for the total MIT, (2) the percent of total repair options selected per level of complexity and (3) the percent of each repair option selected within a level based on the total for that level. Percentages were based upon individual subject totals.

Description. Percent-correct scores and arcsin transformed percent-correct scores (Raus) for the total MIT and for each level of complexity are provided in Tables. 3.31 and 3.32. Tables 3.33 to 3.35 list mean percent-correct scores, standard deviations and ranges for the three sets of repair option data. Figures 3.22 to 3.27 display these results. Standard deviations for repair option and repair option per level are provided in figures separate from the means. The most frequently types of repair options includedfor the test as a whole (Figure 3.23) were no repeat, repeat no change and repeat louder. The pattern of response for complexity level is revealed in Figure 3.32. As level of complexity increases the preferred option shifts such that no repeat decreases in frequency and repeat no change and repeat louder increase.

Statistical Procedures.

Noting the presence of differences among percent-correct scores in complexity levels of the MIT, a one-way repeated measure ANOVA was performed on arcsin transformed percent-correct scores to determine whether differences were significant (Table 3.36). A significant difference due to complexity level was confirmed. To establish which mean pairs differed a Newman-Keuls test was performed. The results are illustrated in Figure 3.29. Level B was

Table 3.31. Mean percent-correct, standard deviations, and ranges for scores of the Multiple Instructions Test presented with communication repair options (N=14).

MIT Level	Mean	S.D.	Range
Level A	67.836	13.414	50-83.3
Level B	88.021	8.413	72.2-100
Level C	75.264	18.597	33.3-95.8
Level D	71.897	15.715	43.3-93.3
Total	75.707	12.508	57.1-91.6

Table 3.32. Mean arcsin transformed percent-correct scores, standard deviations, and ranges of the Multiple Instructions Test presented with communication repair options (N=14).

MIT Level	Mean	S.D.	Range
Level A	65.094	11.578	50-78.75
Level B	85.368	10.824	68.38-104.99
Level C	72.706	17.376	36.35-93.84
Level D	68.951	13.994	44.59-90.10
Total	75.051	13.243	56.36-93.13

Table 3.33. Percent, standard deviations, and ranges of repair options selected based on the total number used by each subject for all items of the Multiple Instructions Test (N=14).

Repair Option	Mean	S.D.	Range
Repair Option 1	44.886	23.3	6.8 -87.5
Repair Option 2	16.75	14.17	0 -38.4
Repair Option 3	17.443	23.314	0 -64.1
Repair Option 4	5.579	12.228	0 -49.6
Repair Option 5	2.85	5.044	0 -19.2
Repair Option 6	5.55	6.869	0 -19.2
Repair Option 7	6.893	12.628	0 -46.1

Table 3.34. Mean, standard deviations, and ranges of the number of repair options selected by level of the Multiple Instructions Test. Calculations are based on the number of repair options a subjects selected for all complexity levels of the MIT (N=14).

MIT Level	Mean	S.D.	Range
Level A	24.736	2.762	17.9 -28.2
Level B	22.071	3.493	15.6 -26.9
Level C	25.557	3.337	18.7 -30.7
Level D	27.264	3.602	23.0 -34.3

Table 3.35. Mean, standard deviations, and ranges of the number of repair options selected within levels of the Multiple Instructions Test. Calculations based on each subjects total number of repair options used within a level of complexity (N=14).

MIT Level	Repair Option	Mean	S.D.	Range
Level A				
	Repair Option 1	50.965	30.736	0-100
	Repair Option 2	14.071	14.648	0 -37.5
	Repair Option 3	17.693	24.727	0 -72.2
	Repair Option 4	6.043	15.193	0 -55.5
	Repair Option 5	3.386	5.71	0 -12.2
	Repair Option 6	2.993	6.045	0 -28.5
	Repair Option 7	4.707	8.71	0 -28.5
Level B				
	Repair Option 1	70.071	20.082	28.5 -100
	Repair Option 2	14.547	12.976	0 -33.3
	Repair Option 3	7.293	13.521	0 -42.8
	Repair Option 4	2.614	6.679	0 -20
	Repair Option 5	1.186	4.437	0 -16.6
	Repair Option 6	2.029	5.157	0 -14.2
	Repair Option 7	2.036	7.617	0 -28.5
Level C				
	Repair Option 1	38.914	28.573	0 -100
	Repair Option 2	15.579	16.345	0 -50.0
	Repair Option 3	21.836	26.953	0 -66.6
	Repair Option 4	3.957	7.965	0 -22.2
	Repair Option 5	5.15	9.123	0 -28.5
	Repair Option 6	5.821	10.03	0 -28.5
	Repair Option 7	9.071	16.42	0 -50
Level D				
	Repair Option 1	27.107	29.554	0 -100
	Repair Option 2	22.964	21.82	0 -57.1
	Repair Option 3	19.743	28.963`	0 -72.7
	Repair Option 4	9.071	18.788	0 -54.5
	Repair Option 5	2.629	7.768	0 -28.5
	Repair Option 6	8.8821	11.201	0 -33.3
	Repair Option 7	9.479	19.134	0 -61.5

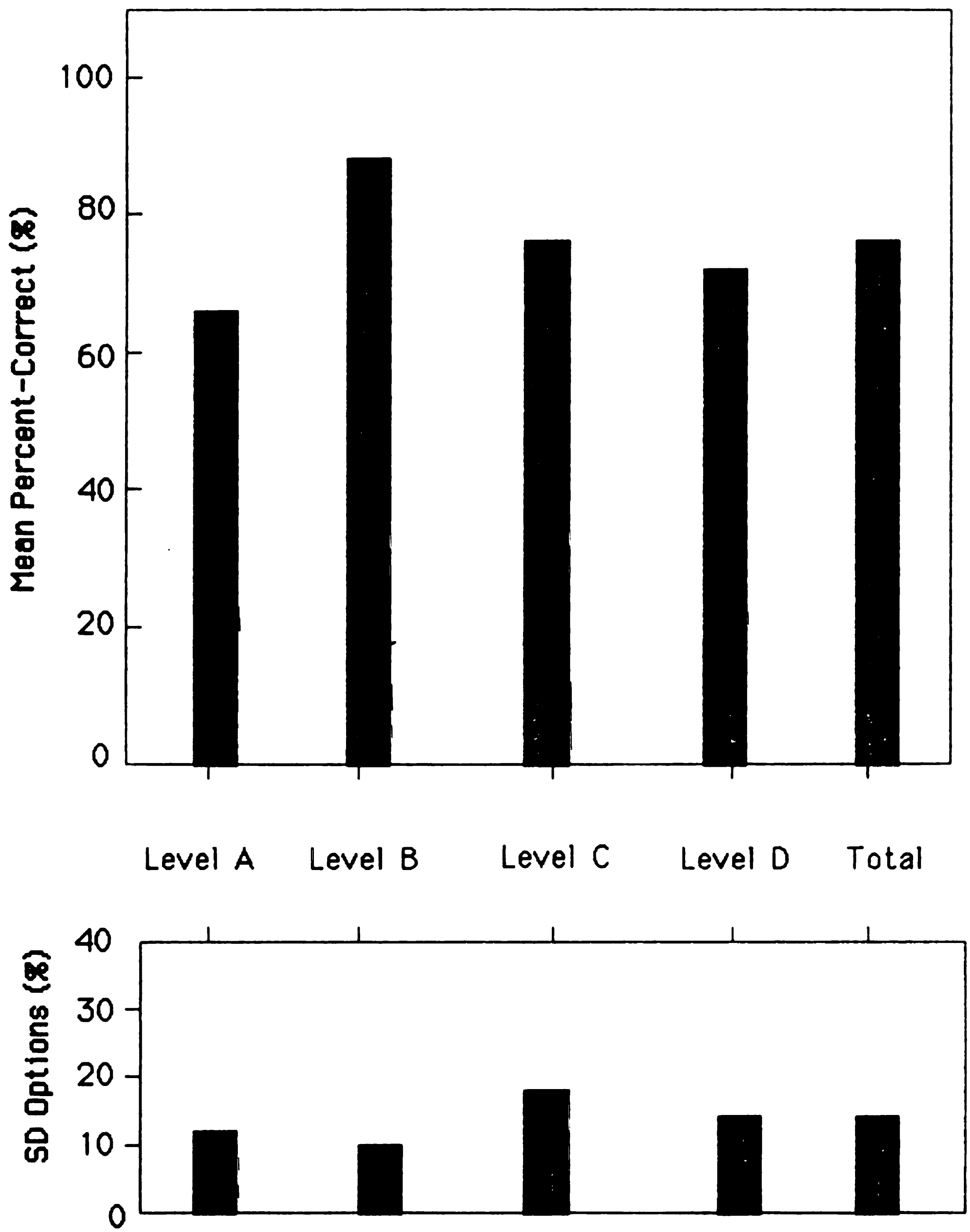


Figure 3.22. Mean and standard deviations percent-correct scores for all levels of test four of the Multiple Instructions Test administered with the option of communication repair.Each bar represents observations of 14 normal-hearing subjects tested monaurally. The lower histogram denotes ±1 standard deviation

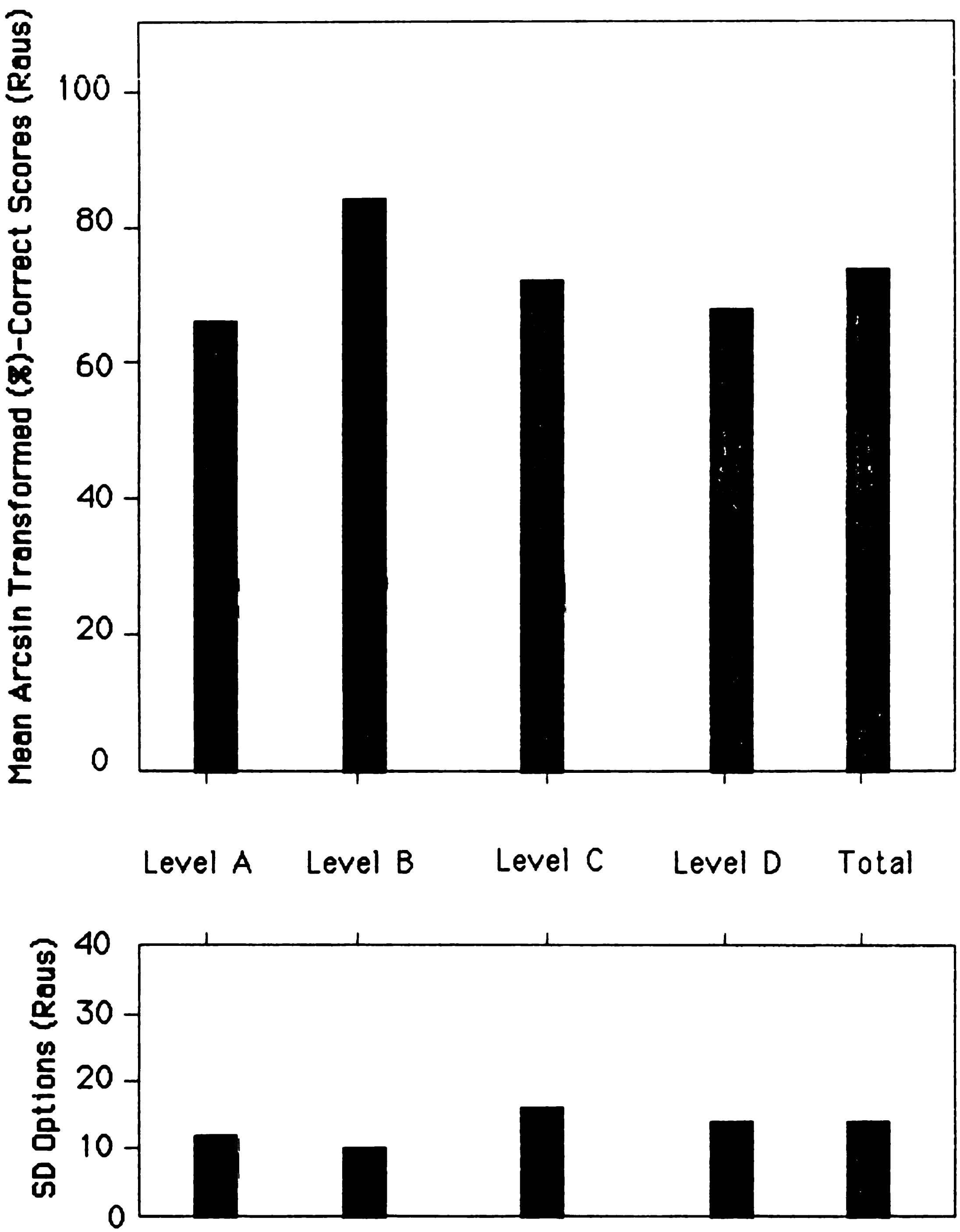


Figure 3.23. Mean arcsin transformed percent-correct scores and standard deviations on all levels of version four of the Multiple Instructions Test presented with communication repair options. Each bar represents observations of 14 normal-hearing subjects tested monaurally. The lower hostogram denotes ±1 standard deviation.

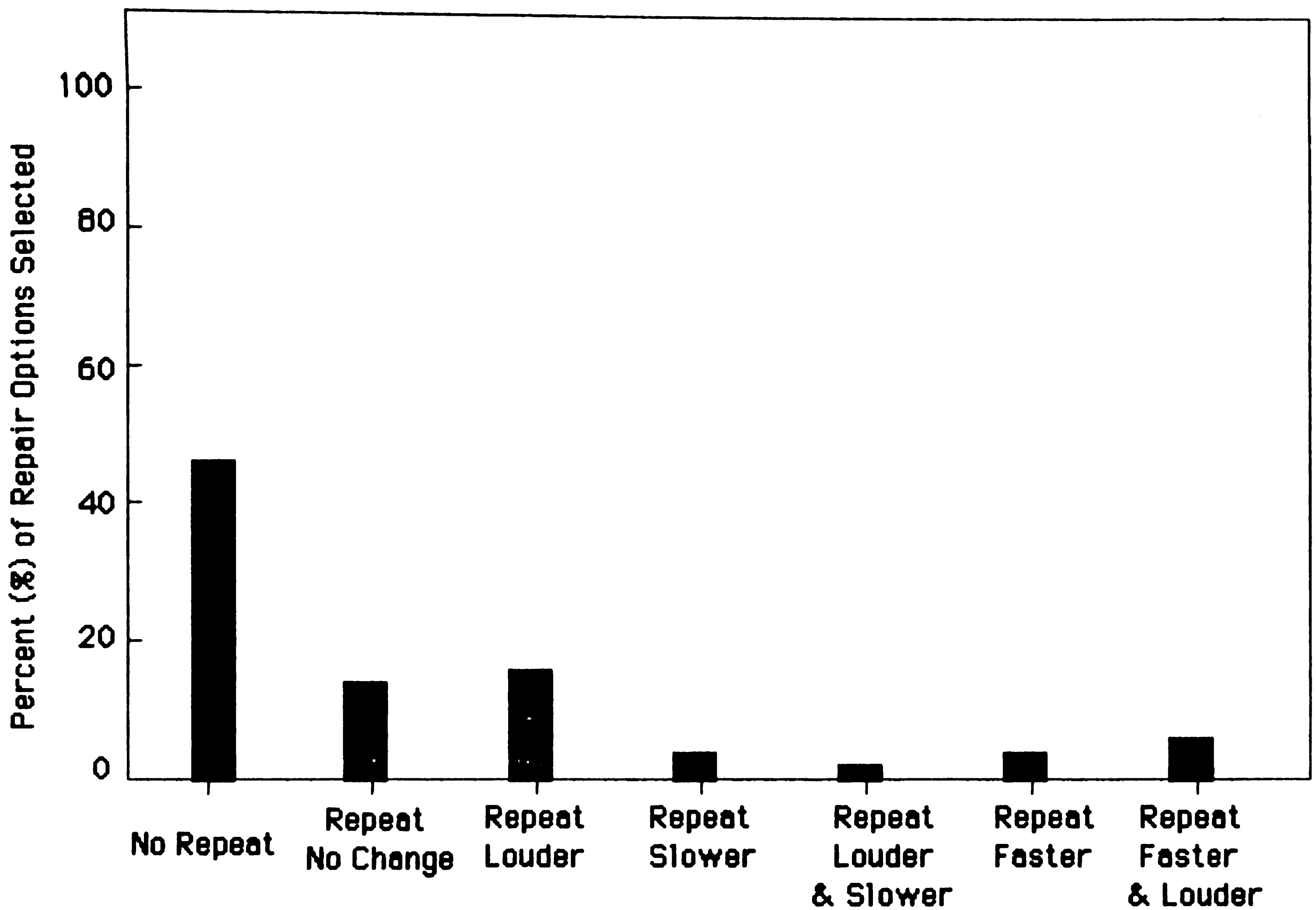


Figure 3.24. Mean percent of repair options selected based on the total number of options selected for all complexity levels of the MIT. Each bar represents observations of 14 normal-hearing subjects tested monaurally.

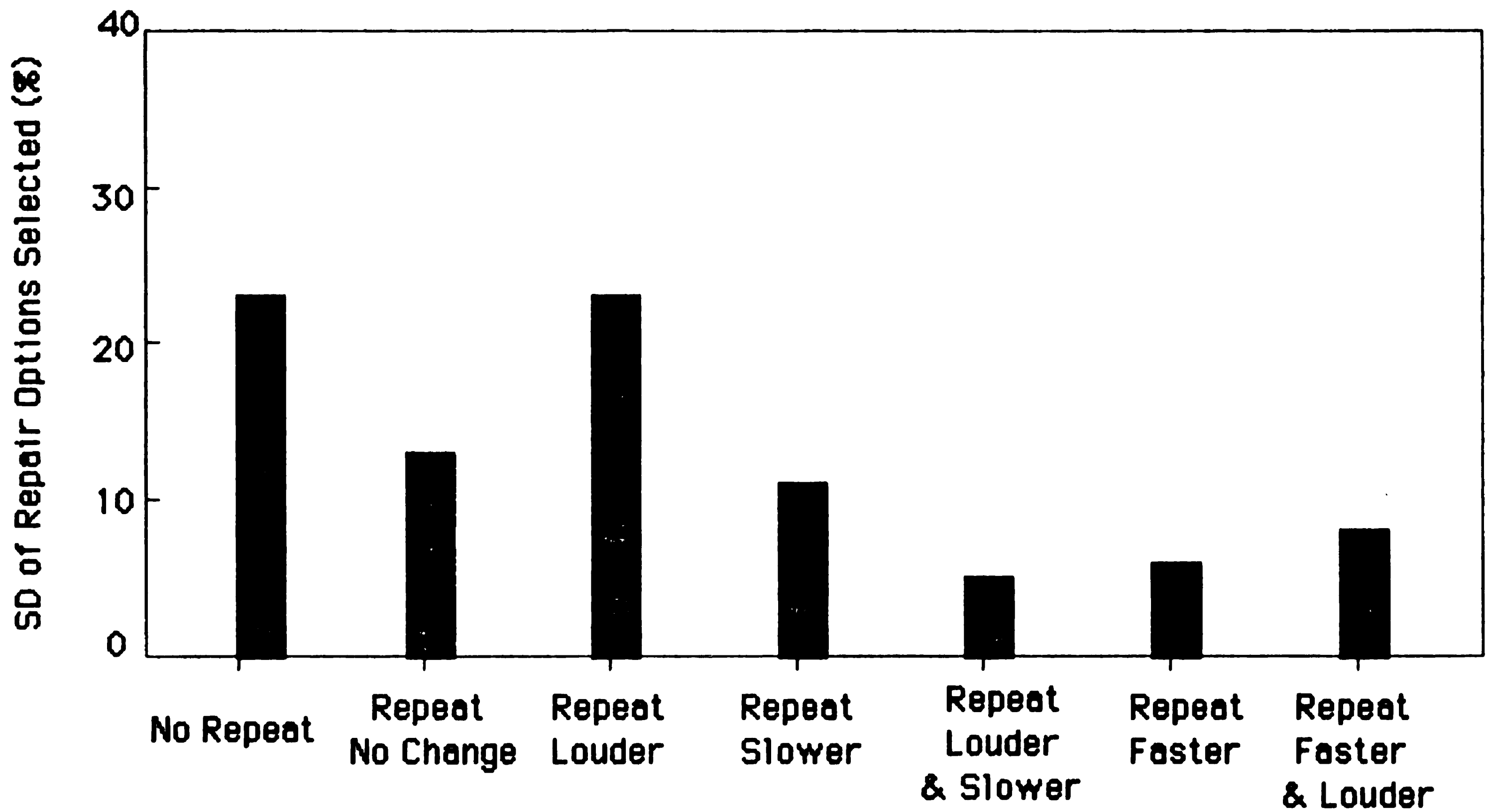


Figure 3.25. Standard deviations for mean percent of repair options selected based on the total number of options selected for all complexity levels of the MIT. Each bar represents observations of 14 normal-hearing subjects and denotes ±1 standard deviation.

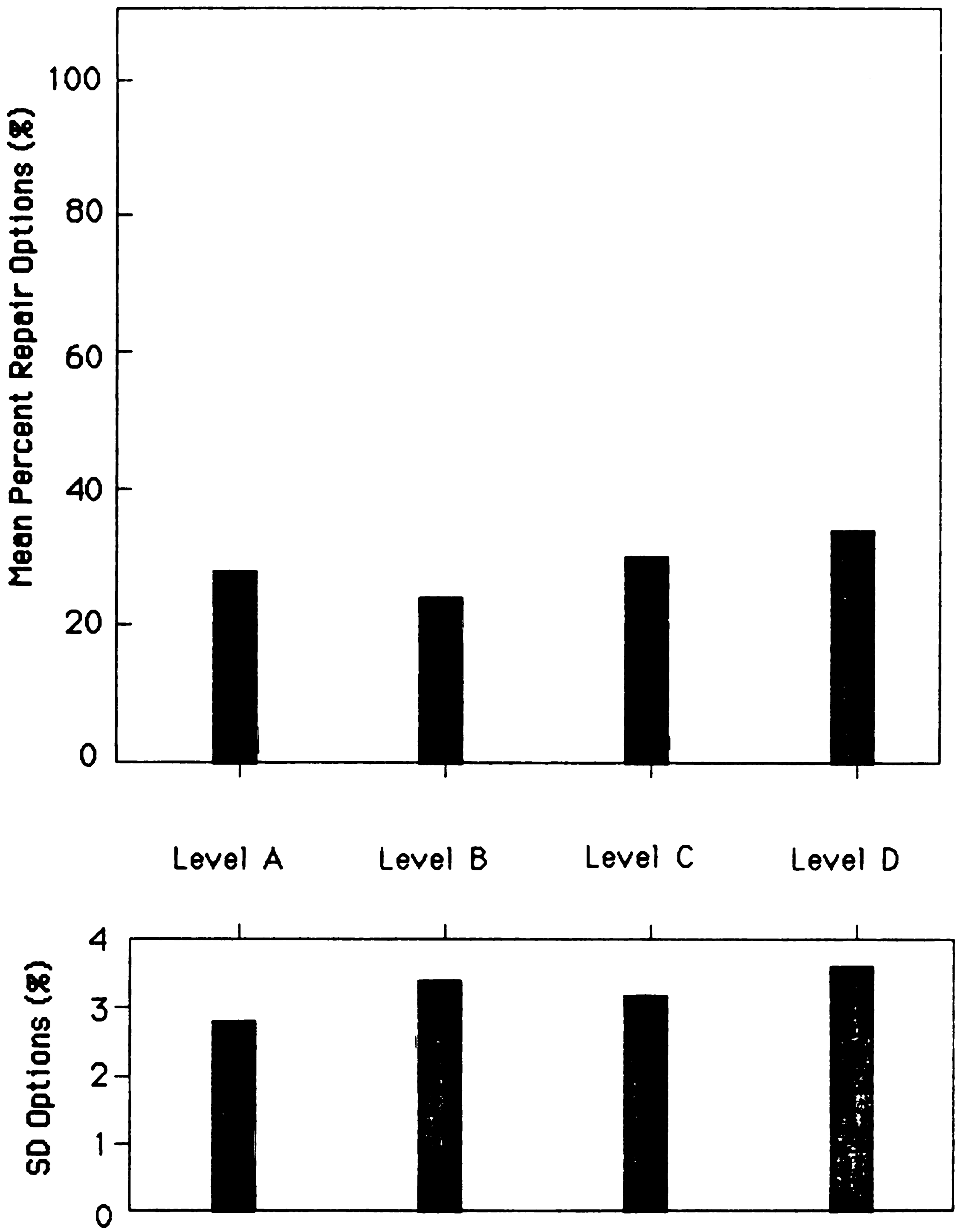


Figure 3.26. Mean percent and standard deviations of options selected at each complexity level based on each subject's total for all levels of the MIT. Each bar represents obsevations of14 normal-hearing subjects. The lower hostogram denotes ±1 standard deviation.

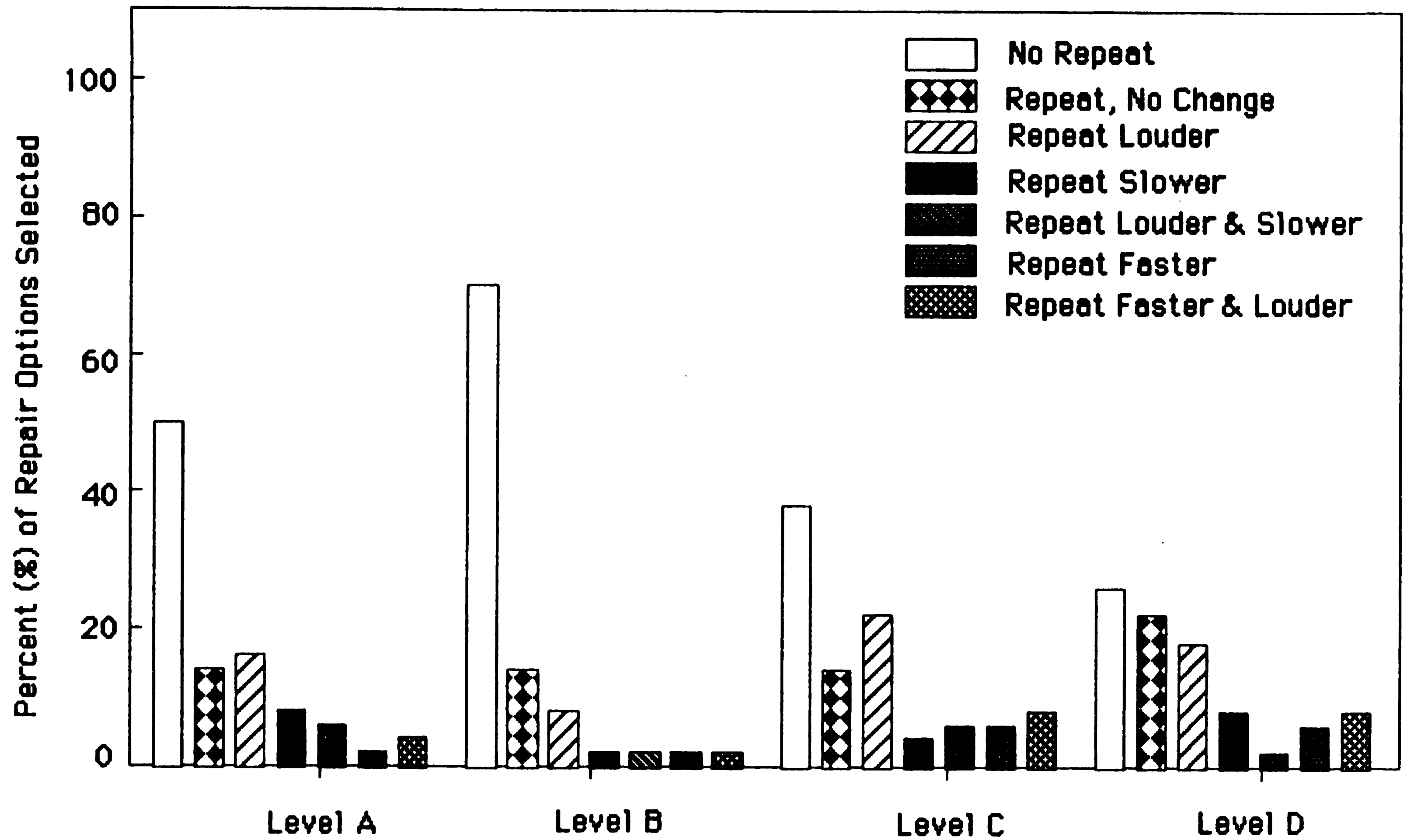


Figure 3.27. Mean percent of repair options selected based on the total number of options for each subject within the complexity level of the MIT. Each bar represents observations of 14 normal-hearing subjects tested monaurally.

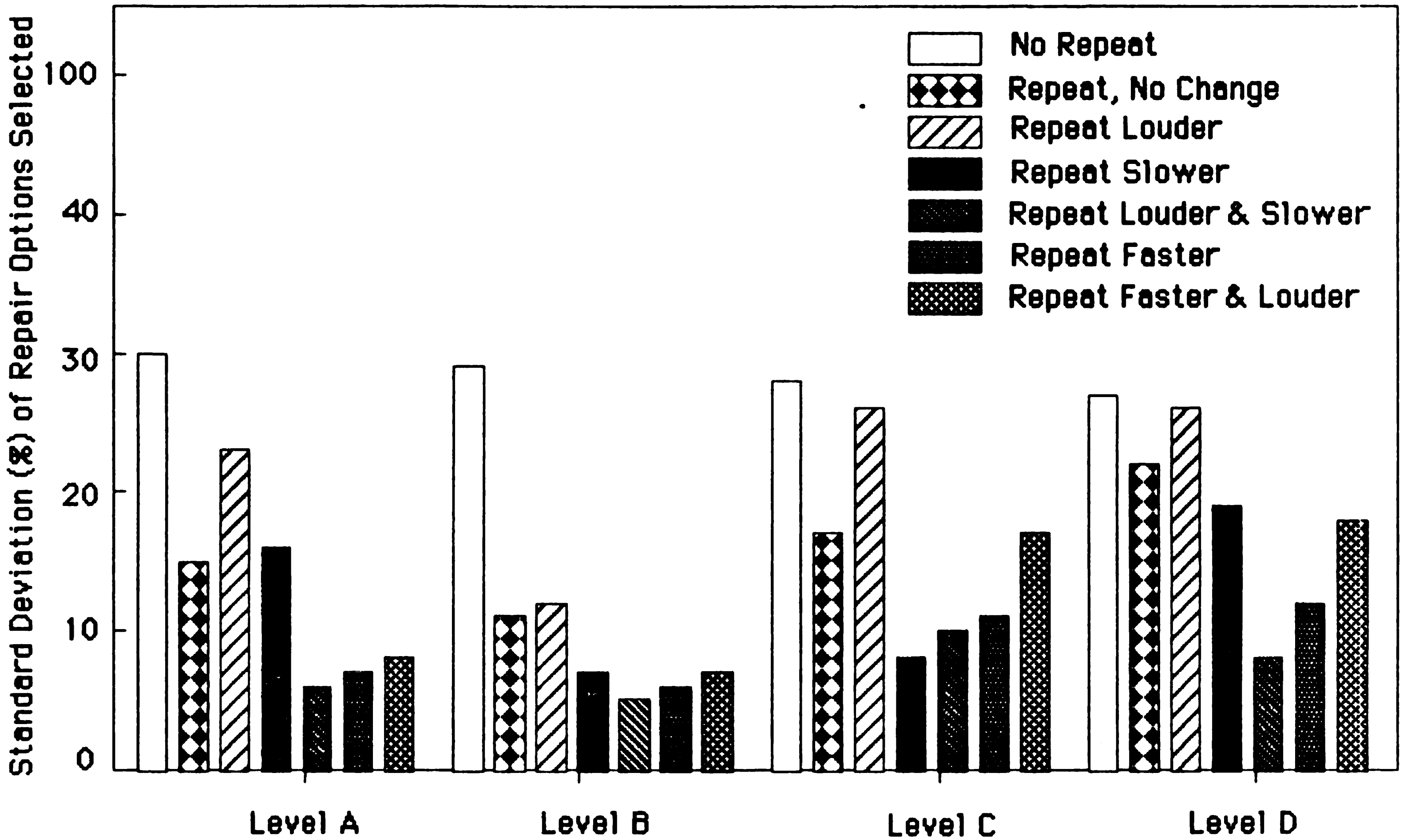


Figure 3.28. Standard deviations for repair options selected within complexity level of the MIT.

Table 3.36. One-way within-subject analysis of variance of total MIT scores as a function of level.

Source	dF	SS	MS	F	p
Subjects	13	6649.628	511.510		
Level	3	3247.368	1982.456	13.675	<.001
Error	39	3087.093	79.156		

Complexity Level	Level A	Level B	Level C	Level D
Mean	65.09			68.95
Percent	65.09		72.70	
Correct		85.36	72.70	68.95

Figure 3.29 Illustration of the results of Newman Keuls' Test of pairwise comparisons of mean percent-correct per complexity level of the MIT. Nonsignificant mean pairs are connected by a solid line.

significantly lower other complexity levels suggesting this that level was less difficult.

A one-way repeated measure ANOVA was performed to assess differences between repetition requests as a function of complexity level

(Table 3. 37) revealed a significant difference. The Newman-Keuls' test revealed only one mean pair (Level B and Level D) differed significantly. Figure 3.30 displays these findings. These results are somewhat consistent with the findings revealing a difference in scores between complexity level however, the percentages used here reflect total repair options selected including "no repeat". Therefore, repetition requests could be balanced out by "no repeat" communication repair requests masking differences between levels. As previously noted, complexity Level B appears to be significantly easier than the others in this version of the MIT. Therefore, a higher number of "no repeat" requests could be expected. Inspection of the percentages does indeed reveal the number of "no repeat" requests in Level B to be higher than any other level (Figure 3.27).

To assess the significance of the two main effects of repair option and MIT complexity level a two-way, repeated measures ANOVA was performed. Table 3.38 provides a summary of the results. No significant difference was seen for the main-effect of complexity. The contradiction between this comparison of MIT level and the one-way discussed previously is thought reflect the different percentages used in

Table 3.37. One-way within-subject analysis of variance of total repair options selected as a function of complexity level of the MIT.

Source	dF	SS	MS	F	p
Subjects	13	2.492	.192		
Complexity Level	3	196.690	65.563	4.496	.008
Error	39	568.715	14.582		

Complexity Level	Level A	Level B	Level C	Level D
Percent Of Repair Options	24.73	22.07	25.55	27.26
	24.73		25.55	
	24.73			27.26

Figure 3.30. Illustration of results of Newman Keuls' Test of pairwise comparisons of the percent of the total repair option selected at each complexity level of the MIT. Nonsignificant mean pairs are connected by solid line.

Table 3.38. Two-way within-subject analysis of variance of Multiple Instructions Test arcsin transformed percent-correct scores with Main Effects of level of complexity and repair option.

Source	dF	SS	MS	F	p
Subjects	13	3.611	.278		
Level	3	.345	.115	.424	.7367
Error	39	10.583	.271		
Repair Option	6	79631.111	13271.85	13.010	<.001
Error	78	79569.83	1020.126		
Level/ Repair Option	18	17976.070	998.671	7.610	<.001
Error	234	30709.097	131.135		

calculations. The one-way ANOVA was based on percentages for the total MIT, while the two way reflects the percent of total repair options selected within level.

The main effect of repair option was found to be significant as was the interaction between level and repair option. These differences were explored further through simple main effects calculations (Table 3.39), the out comes of which revealed an interesting pattern of results. Level at option one (no repeat) and level at option three (repeat louder) were significant. Upon closer inspection of the percentages for Repair option one (no repeat) a decrease can be seen with level of complexity (except at Level B), while request for repetition increase. It can be inferred from these results that an increase in level of task complexity results in an increase in certain types of repair behaviors, specifically requests for repetition and repetition at a higher intensity.

Comparisons Among Experiments

Experiment 1 versus Experiments 2,3,4.

It was previously suggested that selected acoustic properties of speech sources might relate to the accuracy with which a signal is generated in a communication system. Effective creation of speech patterns is dependent upon the underlying rule structure governing that system and in many case these are reflected in phoneme production.

Table 3.39. Simple effects for the variables of Multiple Instructions Test level of item complexity and repair option.

Effect	MSn	DFn	DFe	MSe	F	p
Level at Repair Option 1	4708.147	3	39	338.182	13.922	<.001
Level at Repair Option 2	240.455	3	39	108.512	2.216	.102
Level at Repair Option 3	583.804	3	39	121.529	4.804	.006
Level at Repair Option 4	110.753	3	39	61.758	1.793	.164
Level at Repair Option 5	38.128	3	39	33.352	1.143	.344
Level at Repair Option 6	131.170	3	39	41.579	3.155	.035
Level at Repair Option 7	179.682	3	39	82.773	2.171	.107
Repair Option at Level 1	4113.670	6	78	357.749	11.499	<.001
Repair Option at Level 2	8794.834	6	78	150.562	58.413	<.001
Repair Option at Level 3	2226.464	6	78	58.413	5.709	<.001
Repair Option at Level 4	1132.896	6	78	515.547	2.197	.052

The purpose of the present comparison was to assess the relationship between the acoustic properties of sources and indices of listening performance.

Statistical Procedures. To determine whether characteristics of a sample of vowels were related to indices of semantic precision and task performance, Pearson-product moment correlation coefficients (PPMCCs) were calculated. Summed distance measures for each speech source were used as predictor variables (X) for the dependent variables of word recognition, listening comprehension and task performance scores as measured by SPIN test, Multiple choice test and the Multiple Instructions Test, respectively. In all cases correlation coefficients (r) were low (.3 or less) and therefore, further evaluation of the relationship between variables was not pursued. Thus, the method employed in this study for assessing technical accuracy did not prove to be effective in predicting subject performance.

Implications. The experimental question asked here was whether word-recognition in noise, listening comprehension in noise and complex task performance could be predicted from a single index of technical accuracy based on a sampling of acoustical characteristics. A weak correlation suggests the use of summed vowel distance measures is not in itself an effective means of predicting semantic precision and task performance using speech synthesis systems. While a visual inspection of the distance data reveals individual differences among summed vowel measures, these diferences did not vary in accordance with behavioral tasks data and thus were not a sufficient means of judging the interrelationship between technical accuracy and receiver response.

It is suggested the limited acoustical data gathered here may be both insensitive and insufficient to reflect the underlying rule structures used in the devices. The Hoover et al. (1987) visual inspections of spectograms revealed distinct differences between VOTRAX and Echo both for manner of consonant generation and in the transitions between consonants and vowels. This suggests the need for more elaborate samples of acoustic properties. Spectral and temporal characteristics of consonants, consonant blends, diphthongs and the changes occurring in coarticulation collectively and separately are important and represent further avenues of investigation. Other behavioral tools (e.g. Modified Rhyme Test and Phonetic Specific Sentences) which permit error analysis by phoneme class should be more effective in determining which phonemes to include in a technical analysis.

In addition, the use of means and summed distances appeared to obscure the individual vowel differences associated with each device. DECtalk distance measures for the vowel [ɪ] were large (30.44) in comparison to Smoothtalker (18.85), however, for the vowel [ɑ] the situation was reversed with Smoothtalker demonstrating the greater distance measure. It is also likely that the level of redundancy inherent in sentence and paragraph materials employed in the inherent in sentence and paragraph materials employed in the other experiments may have obscured vowel effects on receiver performance. The additional information present in the stimuli may obscure the differences occurring solely on the basis of vowel structure by allowing the listener to infer words from the surrounding consonant combinations and semantic content.

Experiment 4 versus Experiment 5.

The purpose of this comparison was to determine the effect of communication repair on task performance. Transformed percent-correct scores for the total MIT for Experiment 4 with VOTRAX PSS were compared to those obtained in Experiment 5. Thus, task performance without the opportunity for communication repair was compared to task performance with the opportunity for communication repair.

Description. Table 3.40 and 3.41 give the percent-correct and transformed percent-correct means, standard deviations and ranges for the two conditions of the MIT. Figure 3.31 and 3.32 provide histograms illustrating this data. Calculations are based on twelve subjects (the scores of the last two subjects of Experiment 5 were not included to simplify analysis).

Statistical Procedures. To assess effect of repair option, one-way ANOVA was performed (Table 3.42). A significant difference was found suggesting that the option for communication repair improves task performance.

Table 3.40. Mean percent-correct scores, standard deviations, and ranges of the Multiple Instructions Test presented with (Exp. 5) and without (Exp. 4) the option of communication repair (N=12).

	Mean	S.D.	Range
Experiment 4	67.275	7.163	55.9-79
Experiment 5	75.707	12.508	43.3-93.3

Table 3.41. Mean arcsin transformed percent-correct scores, standard deviations, and ranges of the Multiple Instructions Test presented with (Exp. 5) and without (Exp. 4) the option of communication repair.

	Mean	S.D.	Range
Experiment 4	65.899	6.894	55.28-77.48
Experiment 5	75.051	13.243	56.36-93.13

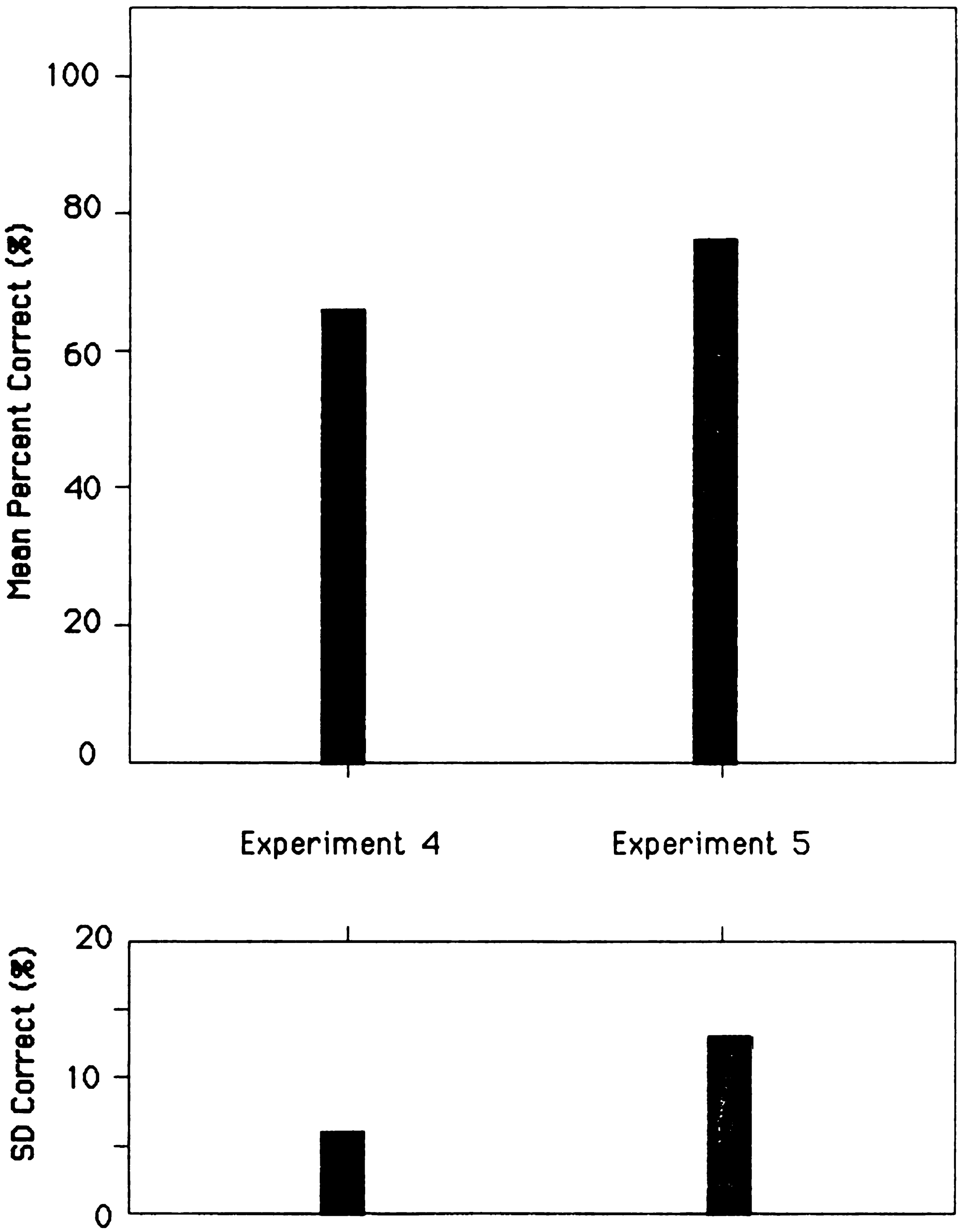


Figure 3.31. Mean and standard deviations of percent-correct scores and standard deviations for the Multiple Instructions Test presented with (Exp. 4) and without (Exp. 5) communication repair options. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

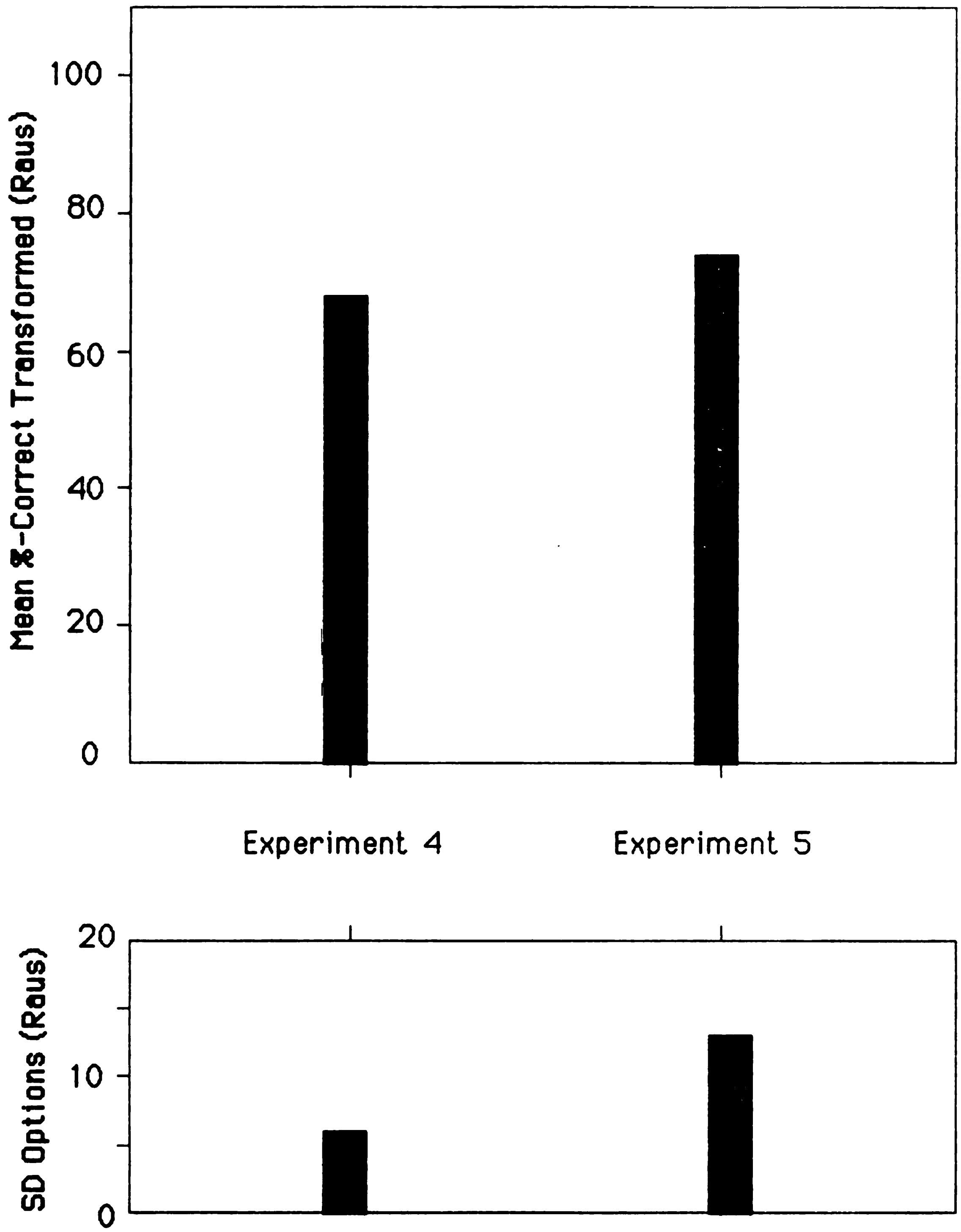


Figure 3.32. Mean and standard deviations of transformed percent-correct scores and standard deviations for the Multiple Instructions Test presented with (Exp. 4) and without (Exp. 5) communication repair options. Each bar represents observations of 12 normal-hearing subjects tested monaurally.

Table 3.42. One-way between-subject analysis of variance of Multiple Insructions Test arcsin transformed percent-correct scores as a function of communication repair.

Source	dF	SS	MS	F	p
Subjects	11	1694.659	154.060		
Test Score	1	267.463	267.467	5.139	.044
Error	11	572.532	52.048		

<u>Implications.</u> The experimental question associated with Experiments 4 and 5 asked is whether the availability of listener-invoked feedback influences performance in complex tasks. The results indicate the option of communication repair improves receiver task performance. These results are consistent with those seen in studies of human`interaction (Leavitt and Mueller, 1951). In addition, the repair options of repeat no change and repeat louder were selected significantly more often than any other repetition option. It is interesting to note that despite the instruction to be as accurate as possible, eight subjects never requested more than two repetitions per item. Yet subject comments following the task suggested that many were dissatisfied with their performance. Thus although communication repair improves performance repetition options alone were often not sufficient to promote confidence in task completion with this speech source. By the structure of the experimental design, all subjects would have requested at least 24 communication repair options (one for each item presented) with a maximum of 72. The actual range was from 24 to 39. It is possible that some subjects perception of synthesized speech was such that they were unaware of errors even with a repetition. For example, four subjects consistently heard the phrase "underline" as "girl in" and drew figures in the corresponding graphics. This was a persistent response through out the task for these subjects. Repetition of the same item or of this instruction within a different command

did not change the response. It would be interesting to discover whether subjects' perception of that instruction change with information about response accuracy and how many repetitions would be required to correct it. It is also possible that some subjects declined to expend the extra effort involved in requesting repair without some indication it was effective.

Implications For System Design

Experiments 4 and 5 support the value of including a feedback option in the design of communication systems, and of recognizing the patterns of response elicited by listeners in overcoming comprehension difficulties. If it is necessary to limit the number of repair options provided to subjects the most logical choices (based on the findings of the present study) would be "repeat" and "repeat louder".

It was shown in Experiment 2 that linguistic predictability also plays a role in intelligibility. Therefore, the inclusion of an option for "repeat enhanced" or "rephrasal" should be investigated. Such an option might allow the user to select from an appropriate list of adjectives to invoke the use of a thesaurus for word substitution.

Clinical Applications

In addition to system design considerations, the differences based upon linguistic predictability should be considered in relation to clinical applications. When clients are being trained to use speech synthesis devices for personal communication the importance of message redundancy should be emphasized. Training should include practice in constructing sentences with sufficient contextual information for listeners to infer key words. Practice in smooth topic transition also should be incorporated. Results of the Revised SPIN test suggest it may be useful as a standardized technique for evaluating speech synthesis systems. The clearly systematic progression of scores seen for both full and half-lists of the SPIN makes it possible to rank the devices on the basis of word recognition scores. Rank ordering is useful in several circumstances: (1) provides a method of comparison for consumers and professionals; (2) when the same device is evaluated at different points in time with either similar or different populations, it provides a framework for comparison; (3) it promotes assessment of effects associated with changes in a device or systems. In addition, relative word recognition scores can be applied to cost benefit decisions about system selection for normal and handicapped populations.

Future Research

The absence of a significant finding in Experiment 1 emphasized the need to learn more about the contribution of acoustic phonetic characteristics to the intelligibility of synthesized speech. Technical analysis of such devices should be expanded to include a larger set of phonemes as well as transition between phonemes. As previously suggested, the MRT and PSS are logical alternative behavioral measures of perception.

Experiment 2 revealed the SPIN as an effective tool for differentiating between speech sources with normal hearing adults. Researchers are beginning to investigate differences in word-recognition of synthesized speech in different populations and age groups (i.e. Mirenda & Beukelman, 1988). An extension of the present study would investigate differences among sources for word recognition as a function of age or degree and type of hearing loss. It also may be possible to generate a "redundancy curve" for speech sources based on sentences which are designed to include varying amounts of information from high to low-predictability. This would provide information about the amount of redundancy necessary for a speech synthesis system to reach a particular level of performance.

Although some significant differences were seen between speech sources and complexity levels in Experiment 4, it might be possible to make the task more sensitive. This could be accomplished by investigating a wider range of vocabulary and restricting the number of times certain key words are presented in and between versions of the test. A simple means of reducing redundancy would be to include more letter commands such as "Place the k in the box". The types of errors seen during error counts suggested these items were more difficult and might provide a more effective means of isolating specific speech source differences.

Designing tasks and performance criterion consistent with those required in industrial or educational settings also are logical extensions of this research. As previously noted, these speech synthesizers tended to group differently depending on the evaluation approach employed. The SPIN resulted in distinct differences among all sources, listening comprehension allowed the grouping of low quality sources and all others, and the MIT permitted low quality synthesizers, mid-to high quality synthesizers and human speech. Further investigation using different tasks (but which maintain a consistent key word vocabulary) might provide more insight on the processing demands associated with speech synthesis devices based on task and source interaction.

Further investigations of the effect of communication repair on task performance also are indicated. These could include studies of the effect of feedback to the subject regarding the correctness of an item and the number of repetitions required to reach 100% accuracy (or some lower asymptotic level) with particular synthesizers. Do patterns of repair options change among speech sources? Do subjects establish a consistent used with all such devices? Answering such questions would provide information about the devices as well as about receiver strategies in communication systems.

Chapter IV

Summary and Conclusion

Introduction

Background

An effective means of organizing the concerns associated with the evaluation of synthesized speech is a communication model. Shannon and Weaver (1949) suggest a triad of issues described as, (1) technical accuracy, (2) semantic precision and (3) task performance. The extent to which researchers have addressed the components of the communication process is varied. Most prior research has concentrated on receptive intelligibility, usually through the use of word or sentence recognition (i.e. Chial, 1973; Greene, Manous & Pisoni, 1984; and listening comprehension tests (i.e. Schwab, Nusbaum & Pisoni, 1985). This approach provides information limited to the semantic accuracy of speech synthesis systems. A few studies have attempted to compare synthesizers on the basis of the quality of the speech and explored the relationship between intelligibility and perceived naturalness. To date, no attempts focused on technical precision and complex task performance. An additional concern is the role of communication repair in human machine interaction. Studies have been reported about the types of repair options selected during performance of a task directed by a speech synthesizer.

Purpose

This study was designed to parallel the components of the Shannon-Weaver model of communication systems. The goal was to:

(1) evaluate a group of speech synthesizers as the three levels of technical accuracy, semantic precision and task performance;

(2) to apply a combination of behavioral techniques used at the same level of the communication process to determine whether different rankings of systems occur as a result of different approaches; and

(3) to obtain initial data on the role of communication repair systems employing synthetic speech.

Experimental Design

Subjects

A total of fifty normal-hearing adult subjects were used in four of the five experiments in this study. Each subject displayed normal sensitivity for pure tones and normal middle ear function. Twelve subjects were randomly assigned to Experiment 2,3 and 4 and fourteen were assigned to Experiment 5.

Speech Sources.

Five devices representing a range of contemporary speech synthesis technology were used in the all experiments reported here These included; (1) DECtalk (Digital Equipment Corporation, 1983), (2) Commodore Amiga, (3) VOTRAX PSS (VOTRAX Incorporated, 1982), (4) Smoothtalker for use with Apple Macintosh (First Byte, 1985), and (5) Echo II +, (Street Electronics, 1984). In addition, a male talker of general American dialect was used as a human control.

Stimulus Recording Methods

Recordings were made on AMPEX 632 tape by connecting the output of the synthesizer to an audio mixer (TEAC, Model MB-20), then to a cassette recorder (JVC, Model KD-15). Stimulus materials were generated using the text-to-speech systems wtih speech sources adjusted for male speech Default conditions for pitch,rate and inflection were used if they existed. Recordings for the human control were made in a sound chamber using a microphone (Electro-Voice, Model RE-15) to mouth distance of 20 to 30 cm. The talker was instructed to use "habitual" inflection, pitch, rate and linguistic emphasis. Successive repetitions of the the level set sentence "He tacked the tip tap top of the teep with toop" was used to set recording levels of all speech sources to a VU meter range of -3 to 0dB. Level calibration tones placed at the beginning of each experimental tape (Experiments 2 - 5) were related to program material such that the calibration tone had an L_{eq} of 6 dB greater than the associated material.

Experiment 1: Acoustic Phonetic Analysis

Stimuli and procedures. Fundamental frequency and formant frequency measures were used to compute spectral loci in three dimensional space for six vowels [i,ɪ,ɛ,æ,ɑ,ʌ] generated by the six speech sources. Formulas used to identify loci were those suggested by Miller (1984).

Spectral loci were also computed for a human model from measurements of vowel spectra made by Peterson and Barney (1952). Distance measures were calculated using a derivation of the Phythagorean Theorem. Spectral loci distance, (relative to the human model) were derived by summing distance measures for individual vowels for each speech source. Findings. Vowel spectrum shapes distance measures as defined by Miller (1984) do not appear to be related to:

(1) word recognition in noise

(2) listening comprehension in noise or

(3) complex task performance.

Experiment 2: Word Recognition

Stimuli and procedures. Semantic accuracy was evaluated using word recognition in sentences. The Revised Speech Intelligibility in Noise test was presented monaurally to subjects in the presence of a twelve-talker-babble noise (+8 dB S/B) by six speech synthesizers and a human control. Performance was measured by mean percent-correct scores for SPIN full-list and subtests (high and low-predictability key words). Thus, the dependent variable of word recognition was investigated on two levels of effect, speech source and linguistic predictability. Arcsin transforms of percent-correct scores (Studebaker, 1985) were analyzed for significant differences as a function of speech source of full-list, linguistic predictability and for speech source - linguistic predictability interaction.

Findings. The findings of Experiment 2 are as follows:

(1) Mean arcsin transformed percent-correct SPIN full-lists scores differed as a function of speech source indicating word recognition was poorer for all speech synthesizers than for the human control.

(2) Mean arcsin transformed percent-correct SPIN subtest scores differed as a function of the linguistic predictability of key words. Low-predictability key word scores were lower than high-predictability key word scores for all speech sources.

(3) Mean arcsin transformed percent-correct SPIN subtest scores differed as a function of the interaction between speech source and linguistic predictability.

Experiment 3: Listening Comprehension

Stimuli and procedures. Semantic accuracy was evaluated using listening comprehension as measured with multiple-choice comprehension tests based upon a set of factual passages. Subjects listened monaurally to passages presented in the presence of a twelve-voice-babble noise (+ 8 dB S/B) proffered by the six speech sources. Subject performance was measured by percent-correct comprehension scores of the tests and the time to test completion in seconds. Transformed scores and test completion time were analyzed for significant differences as a function of speech source.

Findings. The findings of Experiment 3 are as follows:

(1) Mean arcsin transformed percent-correct multiple-choice test scores differed as a function of speech source.

(2) Mean multiple choice test completion time did not differ as a function of speech source.

(3) Although test completion time did not produce signficant differences between speech sources, when

combined with accuracy (%-correct) descriptive differences were noted.

Experiment 4: Task performance

Stimuli and procedures. Task performance was assessed using completion of oral instructions as measured by the Multiple Instructions Test presented monaurally in the presence of a twelve-voice babble noise (+10 dB S/B). Subject performance was assessed using percent-correct scores for the total MIT and for each of four levels of complexity (A to D), total test completion time and item completion time for each level of complexity. Thus, the variable of oral instructions completion was investigated on two levels; speech source and level of task complexity. The variable of item completion time also was investigated in terms of speech source and level of task complexity. Arcsin transforms were performed on the percent-correct scores. Transformed scores and test completion time were analyzed for significant differences as a function of speech source.

Findings.

The findings of Experiment 4 are as follows:

(1) Mean arcsin transformed percent-correct scores differed as a function of speech source.

(2) Mean arcsin transformed percent-correct scores differed as a function of MIT complexity level.

(3.) Mean arcsin transformed percent-correct scores did not differ as a function of the interaction between speech source and MIT complexity level.

(4) Mean MIT item completion time did not differ as a function of speech source.

(5) Mean MIT item completion time differed significantly as a function of complexity.

(6) Mean item completion time differed as a function of the interaction between speech source and level of complexity.

(7) Although item completion time did not prove to be effective in differentiating between speech sources descriptive differences were noted by time and accuracy were combined as a measure of efficiency.

Experiment 5. Task performance with communication repair

Stimuli and Procedures. Communication repair was assessed by evaluating the types and frequency of repair options selected during an oral instructions task. The Multiple Instructions Test was presented monaurally in the presence of a twelve-voice babble (+10 dB S/B) using a single speech synthesizer (VOTRAX PSS). Subjects could request up to three repetitions of individual test items using one of seven repair options (no repeat, repeat no change, repeat louder, repeat slower, repeat louder and slower, repeat faster, repeat faster and louder). Performance was measured by the number of repair options requested, which were then converted to percentages. The percentages were analyzed for significant differences as a function of type of repair option requested, total number requested per MIT level of complexity, type of repair option requested within a level, types of repair options requested between levels and the interaction of type of repair options and level of complexity.

Findings. The findings of Experiment 5 are as follows:

(1) Mean percentages differed as a function of types of repair option requested.

(2) Mean total repair options differed as a function of complexity levels.

(3) Mean percentages differed as a function of type of repair options requested within a level of complexity.

(4) Mean percentages differed as a function of interaction between repair option types within and level of complexity.

Experiment 4 versus Experiment 5. Task performance with communication repair.

Stimuli and procedures. Arcsin transformed percent-correct Multiple Instructions Test scores were used from Experiment 4 as a measure of oral instruction performance without benefit of communication repair option. Arcsin transformed percent-correct MIT scores for Experiment 5 were used as measure of oral instruction performance with benefit of communication repair options. The dependent variable of MIT scores was assessed for significant differences as a function of communication repair.

Findings. The findings for the comparison between Experiment 4 and 5 revealed the mean transformed percent-correct scores differed as a function of communication repair.

Conclusions

In addition to the above findings, the results of this study to provide the basis for the following tentative conclusions:

1. The systematic monotonic relationship seen among SPIN scores as a function of speech source suggest that this test

is sensitive to differences among speech sources and will be useful in future comparisons of speech synthesizers.

2. Complex tasks such as the Multiple Instructions Test appear practical for use in studies involving speech synthesizers and in investigations of various aspects of communication systems.

3. Although patterns of significant differences varied among high and mid-quality speech synthesis devices, two speech sources (Smoothtalker and Echo II+) consistently performed more poorly on all behavioral measures.

4. Communication repair appears to have a positive effect on performance in communication systems employing speech synthesizers and, therefore, deserves further examination as a technique for improving human-computer interaction.

APPENDIX A

Appendix A.

INFORMED CONSENT RELEASE

Experiment #_______

ID# ________________ Seq#__________

1. I, ______________________, freely and voluntarily consent to serve as a subject in a scientific study of speech perception conducted by Laura J. Kelly, M. A. working under the supervision of Michael R. Chial, Ph. D..

2. I understand that the purpose of this study is to determine the usefulness of several speech synthesizers which may have clinical applicability.

3. I understand I will not be exposed to any experimental conditions which constitute a threat to my hearing, nor to my physical or psychological well being.

4. In understand the information gathered in this experiment is confidential, that no information uniquely identified with me will be made available to other persons or agencies, and that any publication of the results of this work will maintain my anonymity.

5. I engage in this study freely, without payment to me or from me, and without personal benefit. I understand that I may cease participation in the study at any time.

6. I have had the opportunity to ask questions about the nature and purpose of the study, and I have been provided with a copy of this informed consent form. I understand that upon completion of the study, and at my request, I can obtain additional explanation about the study.

Date_______________ Signed____________________________

AUDIOLOGICAL SCREENING

Subject Identification

Name:______________________ ID#:____________ DOB:__________ Age:________ Sex:_____

Date:______________ Phone:______________ Handedness:____________ Examiner:____________

History

Recent onset of HL	Yes____	No____	Tinnitus	Yes____	No____
Active URI	Yes____	No____	Noise exposure	Yes____	No____
Genetic history of HL	Yes____	No____	Otologic surgery	Yes____	No____
Vertigo	Yes____	No____	Pharm. agents	Yes____	No____

Release Signed?

Test Results

1. Otoscopy: **Minimal cerumen; grossly normal TMs & canals**
2. Pure Tone Air Conduction Thresholds (HTL): **< 15 dB, each frequency, each ear**

Freq:	250	500	1k	2k	3k	4k	6k	3-Hz Ave.	2-Hz Ave.
R									
L									

Audiometer ________

├─ 500, 1k, & 2k Hz ─┤

3. Typanometry (attach form)

Middle Ear Pressure at PEV (daPa or mm of water at 220 Hz) **Between -100/+50mm**
Peak Equivalent Volume (ml or cc): **Between 0.2 and 1.4 ml**
Type Classification (re: Jerger, 1970): **Type A or Ad, unimodal**

	MEP	PEV	Type
R			
L			

Bridge ________

4. Acoustic Reflex Thresholds (ARTs)

Pure tone reflex thresholds (HTL): **Between 70 and 100 dB**

	Freq.	500	1k	2k
IPSI	R			
	L			
CONTRA	R			
	L			

5. Acoustic Reflex Decay (attach form)

Decrease in reflex amplitude for signals at ART +10 dB: **< 50% in 10 sec.**

	Freq.	500	1k
CONTRA	R		
	L		

Criterion Met?
History Yes___ No___
Release Yes___ No___
Otoscopy Yes___ No___
PT AC Thresholds Yes___ No___
Tympanometry Yes___ No___
Acoustic Reflex Threshold Yes___ No___
Acoustic Reflex Decay Yes___ No___

MRC: 10/15/86

APPENDIX B

Appendix B. Speech Source Parameters

DECtalk (Dectalk, 1984. p. 63)

Perfect Paul

Parameter	Default Settings
Speaker sex (0 - 1)	1
Smoothness (0 - 100)	34%
Assertiveness (0 -100)	100%
Average Pitch (30 - 300)	120 Hz
Pitch Range (0 - 250)	100 %
Breathiness (0 - 70)	0 dB
Richness (0 -100)	20 %
Samples in open period (0 -100)	0
Laryngealization (0 - 100)	0%
Head Size (40 - 200)	100%
Cascade format 4 frequency (2000 - 4650)	3300 Hz
Cascade format 4 bandwidth (100 - 2048)	60 Hz
Cascade format 5 frequency (2500 - 4950)	3900 Hz
Cascade format 5 bandwidth (100 - 2048)	130 Hz
Parallel format 4 frequency (2500 - 4950)	3300 Hz
Parallel format 5 frequency (2500 - 4590)	4050 Hz
Gain of frication source (0 - 80)	73 dB
Gain of aspiration source (0 - 80)	70 dB
Gain of voicing (0 - 80)	71 dB
Gain of nasal resonator (0 - 80)	69 dB
Gain of resonator 1 (0 - 80)	72 dB
Gain of resonator 2 (0 - 80)	61 dB
Gain of resonator 3 (0 - 80)	50 dB
Gain of resonator 4 (0 - 80)	59 dB
Gain of resonator 5 (0 - 80)	72 dB
Fo dependent spectral tilt (0 - 100)	35%
Beginning pitch basline fall (50 - 200)	115 Hz
End pitch baseline fall (50 - 200)	100 Hz

Amiga (Knight, 1986, p.131-132)

Parameter	Default setting
Pitch Range (65 - 320)	110
Inflection or monotone	Inflection
Rate (40 - 400)	150
Sex	Male
Voice Quality (5000 - 28000)	22000
Volume (0 - 64)	64

Appendix B. (continued)

	Default Settings
Votrax PSS (Votrax, 1982, p.18-20)	
Rate 1-9 and A-F, 0	Default 4 Slow 6 Fast 3
Inflection 0-7	1
Amplitude 0-F	Default 7 Louder 1
Text Conversion mode 0 basic translation	
Voice mode standard SC01	
Smoothtalker (First Byte, 1984)	
Speed (0 - 9)	5
Pitch (0 - 9)	5
Volume (0 - 9)	5
Tone (treble and bass)	Bass
Sex Male	
Echo II+ (Street Electronics, 1984, p. 19-28)	
Pitch	24
Inflection and monotone	Inflection
Sex	Male
Amplitude (1 - 15)	7
Speed (talkfast and normal)	Normal

Appendix B. (continued)

Program for text-to-speech conversion on Commadore Amiga adapted from Knight (1986).

```
100  OPEN "dfo: file name" For Input as 1
110  If EOF (1) Then 200
120  Input #1, 1$
140  a$= TRANSLATE$ ($)
150  SAY (a$)
160  GOTO 110
200  PRINT "Done"
210  Close #1
```

APPENDIX C

Appendix C. Equipment and tape calibration check.

Table C-1. Sound level measurements of Revised SPIN Test experimental tape generated using a human talker (male).

Voice			Tape Type		Stimuli				
Human Male			Cassette		SPIN				
Mean Diff. 6.1 SD Diff. 0.4	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	109.9	109.4	110.6	110.8	110.3	110.6	110.3	114.3	110.3
Leq	102.7	103.7	104.6	104.4	104.3	104.2	103.9	108.5	104.7
Diff.	7.2	5.7	6.0	6.4	6.0	6.4	6.4	5.8	5.6
High	109.0	110.8	111.3	111.0	111.0	111.0	110.8	115.8	111.3
Low	76.8	76.8	80.3	76.8	79.5	76.8	76.8	76.8	76.8
Peak	123.6	124.0	125.4	125.4	125.3	125.3	125.3	129.5	125.8
SEL	129.5	130.6	131.6	131.4	131.3	131.2	130.9	135.4	131.6
Time	8:09	8:20	8:24	8:22	8:24	8:25	8:24	8:24	8:27

Table C-2. Sound level measurements of Revised SPIN Test experimental tape generated using the DECtalk speech synthesizer.

Voice			Tape Type		Stimuli				
DECtalk			Cassette		SPIN				
Mean Diff. 7.1 SD Diff. .42	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	109.9	109.4	109.4	108.5	108.8	108.5	110.9	109.6	109.4
Leq	103.3	102.5	102.2	101.9	101.3	100.6	104.0	102.3	102.6
Diff.	6.6	7.2	7.2	6.5	7.5	7.9	6.9	7.3	6.8
High	110.5	109.5	109.5	109.8	108.3	108.0	111.3	110.0	109.8
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	123.5	123.0	122.4	122.8	121.8	122.0	124.8	123.3	123.5
SEL	129.4	128.8	128.4	128.2	127.8	126.9	130.3	128.6	12.8
Time	7:04	7:09	7:05	7:05	7:08	7:08	7:07	7:06	7:03

Appendix C. (continued)

Table 3.3. Sound level measurements of experimental tapes of the Revised SPIN test generated by the Amiga speech synthesizer.

Voice Amiga		Tape Type Cassette			Stimuli SPIN				
Mean Diff. SD Diff.	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	101.2	102.9	102.5	102.6	104.3	102.2	102.1	103.1	102.4
Leq	95.7	96.8	96.5	96.6	98.1	96.3	96.6	97.3	96.6
Diff.	5.5	6.2	6.0	6.0	6.2	5.9	5.5	5.8	5.8
High	102.3	102.3	102.5	102.4	104.0	102.0	102.8	107.8	102.5
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	113.9	116.8	114.8	114.5	116.3	114.1	114.6	114.3	114.4
SEL	122.0	123.0	122.9	122.8	124.4	122.5	122.8	123.6	122.8
Time	7:13	7:12	7:15	7:05	7:10	7:05	7:02	7:10	7:09

Table 3.4. Sound level measurements of experimental tapes of the Revised SPIN test generated by the Votrax PSS speech synthesizer.

Voice Votrax PSS		Tape Type Cassette			Stimuli SPIN				
Mean Diff. 6.5 SD Diff. 0.2	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	105.1	104.4	104.6	104.6	104.6	105.1	105.1	104.3	104.4
Leq	98.4	98.0	94.4	98.3	98.4	98.3	98.3	97.9	97.7
Diff.	6.7	6.4	6.2	6.3	6.2	6.8	6.8	6.4	6.7
High	103.5	103.3	103.8	103.8	103.8	103.6	103.5	103.0	102.5
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	117.8	116.5	116.8	116.6	116.6	116.8	116.6	116.6	116.0
SEL	125.4	124.9	125.4	125.2	125.4	125.3	125.3	124.9	124.7
Time	8:31	8:16	8:23	8:17	8:23	8:22	8:22	8:22	8:26

Appendix C. (continued)

Table C-5. Sound level measurements of experimental tapes of the Revised SPIN test generated by the Smoothtalker (on Macintosh) speech synthesizer.

Voice Smoothtalker			Tape Type Cassette			Stimuli SPIN			
Mean Diff. 6.5 SD Diff. 0.5	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	104.1	104.1	104.1	103.9	100.1	99.3	98.5	97.7	97.0
Leq	98.1	97.4	97.4	96.4	92.9	92.6	92.4	91.6	92.1
Diff.	6.0	6.7	6.7	7.5	7.2	6.8	6.1	6.1	5.9
High	103.8	103.3	103.0	102.8	98.5	98.5	98.0	97.8	98.3
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	118.0	117.4	117.1	117.1	112.5	112.8	112.0	111.6	111.8
SEL	124.8	124.1	124.1	123.0	119.5	119.2	119.1	118.3	118.8
Time	7:51	7:49	7:51	7:46	7:48	7:46	7:56	7:53	7:55

Table C-6. Sound level measurements of experimental tapes of the Revised SPIN test generated by the Echo II + speech synthesizer.

Voice Echo II+ (with inflection)			Tape Type Cassette			Stimuli SPIN			
Mean Diff.6.2 SD Diff. .26	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	108.4	108.4	107.9	107.9	108.4	108.3	108.0	108.3	108.3
Leq	101.9	102.6	102.0	101.4	102.3	101.9	101.9	102.3	101.8
Diff.	6.5	5.8	5.9	6.5	6.1	6.4	6.1	6.0	6.5
High	107.3	108.0	107.5	106.8	107.5	107.5	107.5	107.5	107.5
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	119.1	119.5	118.6	118.3	119.1	119.3	118.9	118.6	119.3
SEL	128.4	129.1	128.6	127.9	128.8	128.4	128.4	128.3	128.4
Time	7.32	7:32	7:38	7:32	7:34	7:32	7:31	7:35	7:35

Appendix C. (continued)

Table C-7. Sound level measurements of experimental tapes of the Revised SPIN test twelve-voise babble as recoded on the second channel of the tape.

Voice	Tape Type				Stimuli				
Human	Cassette				Human Babble				
Mean Diff. 1.3 SD Diff. .5	P	1	2	3	4	5	6	7	8
Cal. Tone (slow)	109.8	109.6	110.3	110.1	110.3	110.5	110.3	114.0	110.8
Leq	107.3	108.1	109.1	108.8	108.6	109.1	108.8	113.9	109.6
Diff.	2.5	1.5	1.2	1.3	1.5	1.4	1.5	0.10	1.0
High	109.5	110.0	111.4	111.5	111.8	111.5	111.0	116.8	112.3
Low	76.8	76.8	80.3	76.8	76.8	76.8	76.8	76.8	76.8
Peak	121.8	121.6	121.4	123.3	122.4	123.1	123.3	128.4	124.3
SEL	129.5	132.2	136.2	135.9	135.7	136.1	130.9	140.9	136.6
Time	8:22	8:32	8:36	8:34	8:25	8:31	8:28	8:27	8:34

Human Babble Cassette Tape Tape Cal. Tone 110.5 Leq 109.7 Diff. .8
High 111.3 Low 76.8 Peak 122.4 SEL 136.9 Time 9:02

Appendix C. (continued)

Table C-8. Sound level measurements of experimental recordings of factual passages generated by a human talker (male).

Voice Human		Tape Type Cassette			Stimuli Passages			
Mean Diff. 5.8 SD Diff. .20		P1	1	2	3	4	5	6
Cal. Tone (slow)		117.6	114.0	114.0	108.6	108.5	108.4	108.5
Leq	118.8	107.8	106.8	102.7	102.6	102.9	102.8	
Diff.	5.6	6.2	5.8	5.9	5.9	5.5	5.7	
High	115.8	110.5	110.8	106.3	105.8	106.8	107.1	
Low	78.3	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	129.4	125.3	125.6	121.8	121.5	124.1	122.4	
SEL	122.1	123.5	122.6	118.3	118.8	119.0	119.0	
Time		:34	:38	:38	:37	:42	:40	:42

Table C-9. Sound level measurements of experimental recordings of factual passages generated by the DECtalk speech synthesizer.

Voice DECtalk		Tape Type Cassette			Stimuli Passages			
Mean Diff. 5.7 SD Diff. .43	P1	1	2	3	4	5	6	
Cal. Tone (slow)		117.3	117.1	117.4	117.6	117.4	117.4	116.3
Leq	111.5	112.3	110.8	112.3	112.1	111.5	109.9	
Diff.	5.8	5.8	6.5	5.3	5.3	5.9	5.4	
High	115.0	115.3	113.8	116.0	114.9	113.5	114.5	
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	126.5	126.8	126.1	126.5	126.9	126.6	126.4	
SEL	127.9	128.1	126.8	128.3	128.0	126.8	127.2	
Time	:43	:39	:40	:38	:42	:45	:43	

Appendix C. (continued)

Table C-10. Sound level measurements of experimental recordings of factual passages generated by the Amiga speech synthesizer.

Voice Amiga		Tape Type Cassette			Stimuli Passages			
Mean Diff. 5.6 SD Diff. .34	P1	1	2	3	4	5	6	
Cal. Tone (slow)		113.6	199.6	109.4	109.6	109.3	109.1	109.1
Leq	108.6	103.5	103.7	104.2	103.6	103.5	103.6	
Diff.	5.0	6.1	5.7	5.4	5.7	5.7	5.8	
High	111.3	106.5	106.5	107.4	106.3	106.5	106.5	
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	122.6	118.1	118.0	118.8	118.1	118.4	118.1	
SEL	125.4	120.8	121.1	121.1	120.9	121.1	121.1	
Time	:47	:53	:55	:49	:54	:55	:56	

Table C-11. Sound level measurements of experimental recordings of factual passages generated by the Votrax PSS speech synthesizer.

Voice Votrax PSS		Tape Type Cassette			Stimuli Passages			
Mean Diff. 6.3 SD Diff. .38	P1	1	2	3	4	5	6	
Cal. Tone (slow)		114.3	110.3	110.4	110.5	110.5	110.0	110.3
Leq	107.7	104.4	104.6	140.3	103.8	103.6	103.6	
Diff.	6.6	5.9	5.8	6.2	6.7	6.4	6.7	
High	110.0	107.0	107.0	107.3	106.3	106.5	106.3	
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	123.4	120.6	121.1	120.4	120.0	119.8	120.0	
SEL	125.2	122.1	122.5	122.3	121.9	121.6	121.9	
Time		:57	:60	:61	:62	:64	:63	:69

Appendix C. (continued)

Table C-12. Sound level measurements of experimental recordings of factual passages generated by the Smoothtalker (Macintosh) speech synthesizer.

Voice Smoothtalker	Tape Type Cassette			Stimuli Passages				
Mean Diff. 5.6 SD Diff. .49	P1	1	2	3	4	5	6	
Cal. Tone (slow)		117.8	118.3	117.6	117.6	117.6	117.6	117.9
Leq	112.1	112.7	112.1	112.1	112.4	111.9	112.6	
Diff.	6.7	5.6	5.5	5.5	5.2	5.5	5.3	
High	115.5	116.0	115.8	115.3	115.8	115.3	116.5	
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	129.0	129.1	129.3	129.1	129.6	129.4	129.3	
SEL	129.4	129.9	129.7	129.4	129.8	129.4	130.1	
Time	:53	:53	:58	:54	:53	:57	:58	

Table C-13. Sound level measurements of experimental recordings of factual passages generated by the Echo II + speech synthesizer.

Voice Echo II +	Tape Type Cassette			Stimuli Passages				
Mean Diff. 5.8 SD Diff. 3.0	P1	1	2	3	4	5	6	
Cal. Tone (slow)		120.4	120.8	120.9	121.1	120.9	120.8	121.0
Leq	114.5	115.4	115.4	115.5	114.6	114.6	115.3	
Diff.	5.9	5.7	5.5	5.6	6.3	6.2	5.7	
High	117.3	117.5	120.3	119.5	117.0	117.5	117.8	
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8	
Peak	126.4	127.6	127.4	127.4	126.0	127.8	127.8	
SEL	131.1	131.9	132.5	132.2	131.7	132.1	131.9	
Time	:46	:48	:51	:46	:51	:55	:53	

Appendix C. (continued)

Table C-14. Sound level measurements of experimental tapes for Multiple Instructions Test generated by a human male.

Voice		Tape Type		Stimuli			
Human		Cassette		Multiple Instructions Test			
Mean Diff. 6.0							
SD Diff. .31	1	2	3	4	5	6	8
Cal. Tone (slow)	98.2	92.1	92.3	91.9	91.9	91.4	92.9
Leq	92.4	86.0	86.6	85.5	86.3	85.1	86.6
Diff.	6.2	6.1	5.7	6.4	5.6	6.3	6.3
High	99.5	92.5	92.1	92.8	92.0	91.5	92.8
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	114.1	113.9	111.4	115.5	108.9	109.3	112.0
SEL	117.0	110.8	111.6	110.1	111.3	109.9	111.6
Time	4:55	5:08	5:18	4:54	5:23	5:08	5:11

Table. C-15. Sound level measurements of experimental tapes for Multiple Instructions Test generated by the DECtalk speech synthesizer as a speech source.

Voice		Tape Type		Stimuli			
DECtalk		Cassette		Multiple Instructions Test			
Mean Diff. 5.8							
SD Diff. .20	1	2	3	4	5	6	8
Cal. Tone (slow)	102.9	103.1	102.9	103.3	103.1	103.1	102.6
Leq	96.9	97.6	96.8	97.4	97.4	97.1	96.8
Diff.	6.0	5.5	6.1	5.9	5.7	6.0	5.8
High	103.0	103.0	102.5	102.5	103.5	102.6	103.0
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	114.4	114.9	114.1	115.3	114.8	114.8	114.3
SEL	121.6	122.5	121.6	122.1	122.3	122.0	121.4
Time	5:01	5:16	5:01	5:16	5:16	5:07	5:47

Appendix C. (continued)

Table C-16. Sound level measurements of experimental recordings for the Multiple Instructions Test generated by the Amiga speech synthesizer as a speech source.

Voice Amiga	Tape Type Cassette			Stimuli Multiple Instructions Test			
Mean Diff. 6.3 SD Diff. .17	1	2	3	4	5	6	8
Cal. Tone (slow)	103.8	103.6	103.9	103.9	103.9	103.6	103.4
Leq	97.4	97.4	97.4	97.4	97.6	97.3	97.4
Diff.	6.4	6.5	6.3	6.3	6.5	6.3	6.0
High	102.0	102.0	102.4	102.0	102.0	102.5	102.0
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	114.6	114.9	113.3	113.5	113.5	113.5	113.6
SEL	122.8	123.1	123.1	123.1	123.1	122.8	123.1
Time	5:54	6:22	6:10	6:00	6:18	5:56	6:04

Table C-17. Sound level measurements of experimental recordings of the Multiple Instructions Test generated by the Amiga speech synthesizer as a speech source.

Voice Votrax PSS	Tape Type Cassette			Stimuli Multiple Instructions Test			
Mean Diff. 6.3 SD Diff. .22	1	2	3	4	5	6	8
Cal. Tone (slow)	103.8	103.0	103.4	103.6	102.9	103.0	103.6
Leq	97.3	96.5	97.0	97.3	96.6	97.1	97.1
Diff.	6.5	6.5	6.4	6.3	6.3	5.9	6.5
High	101.8	101.0	101.3	101.3	100.5	101.3	101.3
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	114.6	114.4	113.6	113.8	113.5	113.9	114.0
SEL	123.3	122.9	123.1	123.3	122.8	123.1	123.1
Time	6:44	7:10	6:33	6:45	7:01	6:46	6:46

Appendix C. (continued)

C-18. Sound level measurements of experimental recordings of the Multiple Instructions Test generated by the Smoothtalker(on Macintosh) speech synthesizer.

Voice Smoothtalker	Tape Type Cassette			Stimuli Multiple Instructions Test			
Mean Diff. 5.7 SD Diff. .22	1	2	3	4	5	6	8
Cal. Tone (slow)	103.6	103.6	103.8	103.1	104.4	103.9	103.6
Leq	98.1	97.9	98.0	97.4	98.4	97.9	98.1
Diff.	5.5	5.5	5.8	5.7	6.0	6.0	5.5
High	103.3	103.6	103.0	103.8	102.5	102.8	103.3
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	116.8	116.4	116.9	116.1	116.3	116.3	116.6
SEL	123.6	123.6	123.6	123.1	123.1	123.3	123.8
Time	5:59	6:16	6:10	6:10	6:00	5:56	6:01

Table C-19. Sound level measurements of experimental recordings of the Multiple Instructions Test generated by the Echo II + speech synthesizer.

Voice Echo II +	Tape Type Cassette			Stimuli Multiple Instructins Test			
Mean Diff. 5.8 SD Diff. .39	1	2	3	4	5	6	8
Cal. Tone (slow)	103.8	104.9	106.8	104.0	104.3	102.8	103.3
Leq	97.5	98.9	99.3	98.3	97.8	97.3	97.6
Diff.	6.3	6.0	5.5	5.7	6.5	5.5	5.7
High	101.8	103.3	103.5	102.5	102.3	101.5	102.0
Low	76.8	76.8	76.8	76.8	76.8	76.8	76.8
Peak	113.5	112.3	115.1	112.0	120.8	110.6	110.6
SEL	122.9	124.4	124.8	123.8	123.4	122.7	123.0
Time	5:50	5:00	5:59	5:59	5:02	5:52	5:50

Appendix C. (continued)

Voice

(Cassette) Stimulus	H	V	E	A	D	S	Across Voices
SPIN P	7.2	6.7	6.5	5.5	6.6	6.0	M 6.4 SD 0.5
SPIN 1	5.7	6.4	5.8	6.2	6.9	6.7	M 6.2 SD 0.4
SPIN 2	6.0	6.2	5.9	6.0	7.2	6.7	M 6.3 SD 0.4
SPIN 3	6.4	6.3	6.5	6.0	7.6	7.5	M 6.7 SD 0.6
SPIN 4	6.0	6.2	6.1	6.2	7.5	7.1	M 6.5 SD 0.5
SPIN 5	6.4	6.8	6.4	5.9	7.9	6.8	M 6.7 SD 0.6
SPIN 6	6.4	6.8	6.1	5.5	6.9	6.1	M 6.3 SD 0.4
SPIN 7	5.8	6.4	6.0	5.8	7.3	6.1	M 6.2 SD 0.5
SPIN 8	5.6	6.7	6.5	5.8	6.8	5.9	M 6.2 SD 0.4
Across Lists	M 6.1 SD 0.4	M 6.5 SD 0.2	M 6.2 SD 0.2	M 5.8 SD 0.2	M 7.1 SD 0.4	M 6.5 SD 0.5	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table .C-20 Summary statistics for Leg differences between calibration tones and SPIN experimental cassette recordings.

Appendix C. (continued)

Voice

(Cassette) Stimulus	H	V	E	A	D	S	Across Voices
Passage P1	5.6	6.7	5.9	5.0	5.9	6.7	M 5.9 SD .65
Passage 1	6.2	5.9	5.7	6.1	5.8	5.6	M 5.8 SD .23
Passage 2	5.8	5.8	5.5	5.7	6.5	5.5	M 5.8 SD .36
Passage 3	5.9	6.2	5.6	5.4	5.3	5.5	M 5.6 SD .33
Passage 4	5.9	6.7	6.3	5.7	5.3	5.2	M 5.85 SD .57
Passage 5	5.5	6.4	6.2	5.7	5.9	5.5	M 5.8 SD .37
Passage 6	5.7	6.7	5.7	5.8	5.4	5.3	M 5.7 SD 49
Across Passages	M 5.8 SD .2	M 6.3 SD .38	M 5.8 SD .30	M 5.6 SD .34	M 5.7 SD .43	M 5.6 SD .49	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table. C-21 Summary statistics for Leq differences between calibration tones and passage experimental recordings.

Appendix C. (continued)

Voice

Stimulus	H	V	E	A	D	S	Across Voices
MIT P	6.2	6.5	6.3	6.4	6.0	5.5	M 6.1 SD .36
MIT 2	6.1	6.5	6.0	6.5	5.5	5.5	M 6.0 SD .44
MIT 3	5.7	6.4	5.5	6.3	6.1	5.8	M 5.9 SD .35
MIT 4	6.4	6.3	5.7	6.3	5.9	5.7	M 6.2 SD .32
MIT 5	5.6	6.3	6.5	6.5	5.7	6.0	M 6.1 SD .39
MIT 6	6.3	5.9	5.5	6.3	6.0	6.0	M 6.0 SD .29
MIT 8	6.3	6.5	5.7	6.0	5.8	5.5	M 5.9 SD .37
							M SD
Across Tests	M 6.0 SD .31	M 6.3 SD 2.2	M 5.8 SD .39	M 6.3 SD .17	M 5.8 SD .20	M 5.7 SD .22	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table C-22 Summary statistics for Leq differences between calibration tones and Multiple Instructions Test experimental cassette recordings.

Appendix C. (continued)

Voice

(Cassette) Stimulus	H	V	E	A	D	S	Across Voices
SPIN P	489	511	452	430	428	471	M 463.5 SD 30.2
SPIN 1	500	496	455	432	429	469	M 463.5 SD 27.9
SPIN 2	504	503	458	435	425	471	M 466.0 SD 30.3
SPIN 3	502	497	452	425	425	466	M 461.1 SD 30.7
SPIN 4	504	503	454	430	428	468	M 464.5 SD 30.7
SPIN 5	505	502	452	425	428	466	M 463.0 SD 31.8
SPIN 6	508	502	451	422	427	476	M 463.3 SD 33.7
SPIN 7	504	500	455	430	426	473	M 464.6 SD 30.7
SPIN 8	507	506	455	429	423	475	M 465.0 SD 33.4
Across Lists	M 503 SD 2.9	M 502 SD 4.2	M 453 SD 2.0	M 428 SD 3.7	M 426 SD 1.8	M 470 SD 3.4	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table C-23. Total running time in seconds of experimental recordings of the SPIN test.

Appendix C. (continued)

Voice

(Cassette) Stimulus	H	V	E	A	D	S	Across Voices
Passage P1	34	57	46	47	36	53	M 45.5 SD 9.0
Passage 1	41	60	48	53	39	53	M 49 SD 7.9
Passage 2	38	61	51	53	40	58	M 50.1 SD 9.3
Passage 3	38	62	46	49	38	54	M 46.8 SD 9.8
Passage 4	37	64	51	54	42	53	M 50.1 SD 9.5
Passage 5	42	63	55	55	45	57	M 52.8 SD 7.8
Passage 6	40	69	53	56	43	58	M 53.1 SD 10.5
Across Passages	M 38 SD 2.6	M 62 SD 3.7	M 50 SD 3.4	M 52.4 SD 3.2	M 40.4 SD 3.1	M 55.1 SD 2.4	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table C-24 Total running time in seconds for cassette recordings of passages.

Appendix C. (continued)

Voice

Stimulus	H	V	E	A	D	S	Across Voices
MIT P	290	404	350	351	301	359	M 342.5 SD 41.6
MIT 2	308	430	300	382	316	376	M 352.0 SD 51.9
MIT 3	318	393	359	346	301	370	M 347.8 SD 33.8
MIT 4	294	405	359	360	301	370	M 348.1 SD 42.7
MIT 5	323	421	302	378	316	360	M 350.0 SD 44.9
MIT 6	308	406	352	356	307	356	M 347.5 SD 36.8
MIT 8	311	406	350	364	347	361	M 356.5 SD 30.7
Across Tests	M 307 SD 11.9	M 409 SD 12.2	M 338 SD 26.1	M 362 SD 13.3	M 312 SD 16.5	M 364 SD 7.3	

Voice Key

H	Human Male
V	Votrax PSS
E	Echo II+
A	Amiga
D	DECtalk
S	SmoothTalker

Table C-25. Total running time in seconds of experimental recordings of the Multiple Instructions Test.

Appendix C. (continued)

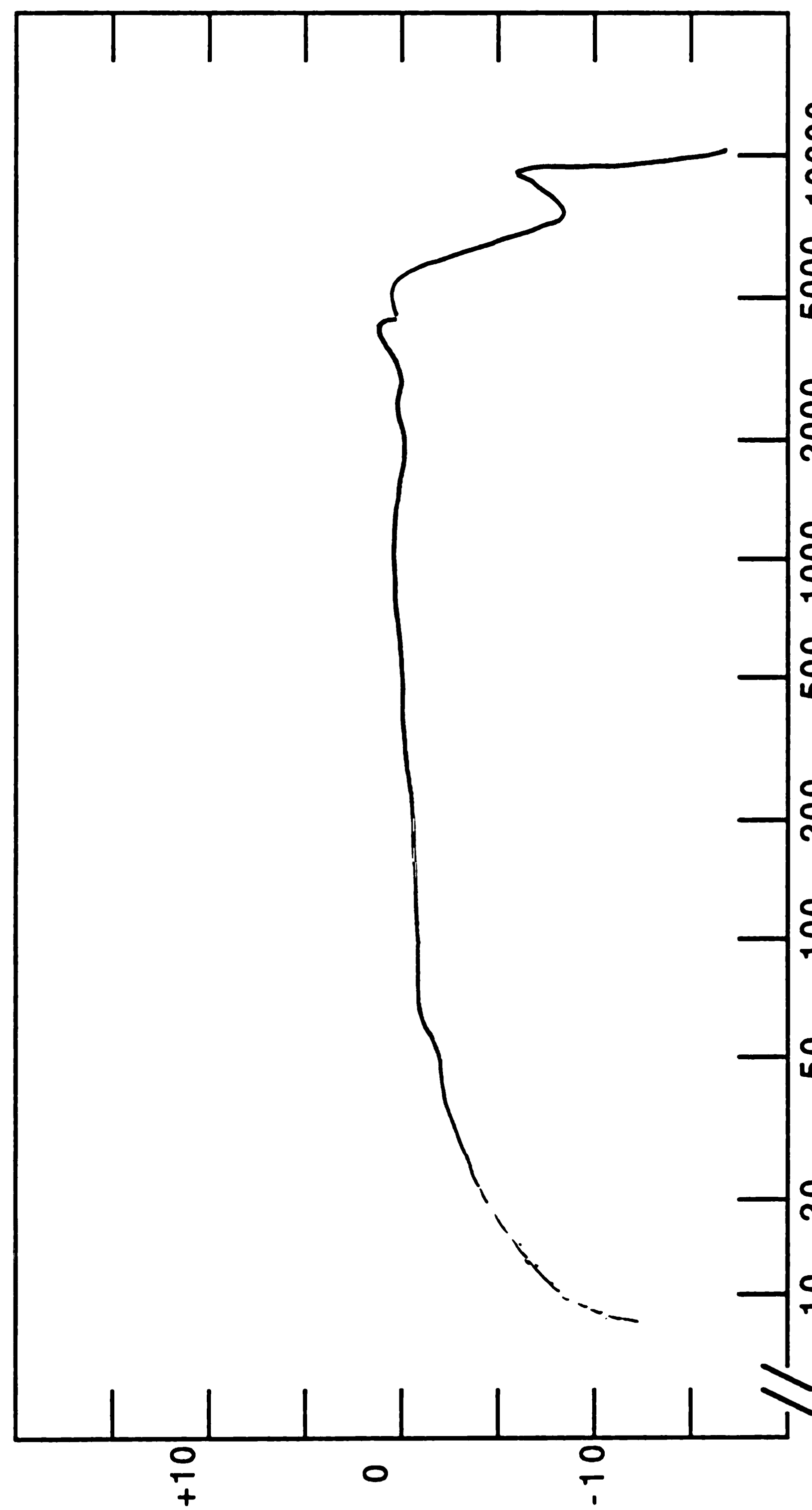


Figure C-1. Frequency response of the experimental earphone (TDH-39)

APPENDIX D

Appendix D. Determination of Revised SPIN Signal-to-Babble Ratios.

INTRODUCTION

A pilot study was necessary to identify a speech-to-babble (S/B) ratio that could be used with all of the speech sources selected for study. The signal-to-babble ratio recommended by Bilger (1983) for use the the revised SPIN test was based upon signals recorded by a human talker. Chial (1973), however, demonstrated that performance-intensity functions may differ for human and synthesizer sources. Previous research has also demonstrated the use of high-quality speech synthesis systems can result in word recognition with low error rates (Greene, Logan, Pisoni, 1986). To reduce the likelihood of a ceiling effect a S/B was sought which would result in a mean full-list SPIN score using high quality synthesized speech of approximately 70 to 90 percent-correct. It was also deemed necessary to monitor the performance of a low quality speech synthesis device to ensure the score did not "bottom out".

The purpose of this pilot study was to find a compromise S/B for which listeners performance with a high quality source (DECtalk) and a low quality source (Echo II +) fell within the linear portion of the performance-intensity function.

SUBJECTS

Seven normal-hearing adult subjects were selected according to previously described criteria for hearing and exposure to synthesis speech.

APPARATUS

The instrumentation employed was that which will be used in the main experiment (see figure 2.3).

PROCEDURE

Subjects were screened and seated in a double walled test booth (IAC, Dimensions 2.54 m x 2.74 m x 1.98 m). Thresholds were obtained using the voice-babble signal presented by monaural earphone (TDH 39 mounted on an MX 41/AR cushion).

Subjects were instructed as follows:

"This is an experiment in which you will hear several sets of sentences. The sentences will come from the earphone on your _______ ear. Your job will be to repeat the last word of each sentence. For example, if your hear "Mrs. Smith did not consider the door," then say "door." It will be hard to hear the sentences because they will be played in the presence of background noise of many people talking at the same time. The noise will come from the same earphone as the sentences. If your are not sure of the last word, feel free to guess.

We will use 8 tests, with 50 sentences. Each test takes a few minutes and will be presented by a different talker at different noise levels. There will be a short break after each test.

Once again, your are to listen to the sentence and repeat the last word you hear. If you are not sure of the word please guess. Any questions? Let's try a practice list."

Each subject was presented with a 30 item practice list composed of 15 items from each speech source. All eight versions of the revised SPIN were used for the experimental sequence. Four S/B ratios were selected +8, +4, 0 and -4 dB HTL. Speech sources were presented in random order. SPIN lists were counter balanced within sources and S/B ratios.

DATA ANALYSIS AND DISCUSSION

Mean percent-correct, standard deviations and ranges were computed for each S/B ratio. As previously discussed, all percent-correct data was transformed to allow for direct comparison in the performance range. Tables D-1 through D-4 summarize data for SPIN full-list and high and low-predictability subtests. Full list SPIN scores ranged from 84.2 % for the DECtalk in the + 8 dB S/B condition to 17.1 % for the Echo II + in the -4 dB S/B condition. High-predictability key word scores ranged from 96.5% for DECtalk (+8 dB S/B) to 17.1 % for Echo II + (-4 dB S/B). Low-predictability key word scores ranged from 72% for DECtalk (+8 dB S/B) to 10.8% for Echo II + (-4 dB S/B). These results are illustrated in Figures D-1 through D-4.

On the basis of these data, a S/B ratio of +8 dB HTL was selected for use in Experiment 2.

Appendix D. (continued)

Table D-1. Mean percent-correct scores, standard deviations, and ranges for the SPIN full-list as a function of signal-to-babble ratio using DECtalk and Echo II + as speech sources (N=7).

Speech Source	Signal-To-Babble Ratio	Mean	S.D.	Range
DECtalk				
	+8	84.2	6.8	76 -94
	+4	79.4	7.2	70 -88
	0	64.5	8.0	54 -76
	-4	35.1	15.9	14 -62
Echo II+				
	+8	37.1	11.7	18 -48
	+4	34.5	13.0	24 -58
	0	28.5	11.4	14 -50
	-4	17.1	14.9	6 -50

Appendix D. (continued)

Table D-2. Mean arcsin transformed percent-correct scores, standard deviations, and ranges for the SPIN full-list as a function of signal-to-babble ratio using DECtalk and Echo II+ as speech sources (N=7).

Speech Source	Signal-To-Babble Ratio	Mean	S.D.	Range
DECtalk				
	+8	84.9	8.9	74.8 -98.4
	+4	79.0	8.1	68.7 -89.0
	0	63.6	7.7	53.6 -74.8
	-4	35.6	15.5	13.5 -61.0
Echo II+				
	+8	37.7	11.5	18.0 -48.1
	+4	35.3	12.3	24.1 -57.3
	0	29.3	11.4	13.5 -50.0
	-4	15.8	1459	1.5 -50.0

Appendix D.

Table D-3. High-predictability and low-predictability SPIN test mean percent-correct scores, standard deviations, and ranges as a function of signal-to-babble ratio using DECtalk and Echo II+ as speech sources (N=7).

Speech Source / Signal-To-Babble Ratio	Mean	S.D.	Range
DECtalk			
High-predictability			
+8	96.5	3.5	92 -100
+4	88.5	8.1	76 -100
0	85.1	9.1	72 -96
-4	41.1	22.3	34 -68
Low-predictability			
+8	72	19.8	60 -80
+4	66.8	9.4	56 -80
0	44	8.6	32 -60
-4	29.1	13.0	20 -56
Echo II+			
High-predictability			
+8	38.4	17.8	5 -56
+4	36.5	13.5	24 -64
0	35.4	9.6	24 -52
-4	17.1	10.2	24 -248
Low-predictability			
+8	23.8	14.4	7 -40
+4	25.7	13.4	12 -52
0	21.7	15.6	4 -48
-4	10.8	9.4	4 -28

Appendix D.(continued).

Table D-4. High-predictability and low-predictability SPIN test mean arcsin transformed percent-correct scores, standard deviations and ranges as a function of signal-to-babble ratio using DECtalk and Echo II+ as speech sources (N=7).

Speech Source Signal-To-Babble Ratio	Mean	S.D.	Range
DECtalk			
High-predictability			
+8	104.3	9.2	93.8 -113.8
+4	91.0	13.1	74.3 -113.8
0	85.6	11.4	70.3 -100.7
-4	40.5	23.0	-.77 -66.44
Low-predictability			
+8	70.9	11.1	59 -88.1
+4	65.6	9.0	55 -78.6
0	44.5	7.8	33.5 -59
-4	30.3	12.3	21.3 -55.3
Echo II+			
High-predictability			
+8	38.3	18.7	1.1 -55.3
+4	36.5	12.4	25.6 -62.68
0	35.5	8.9	25.6 -51.7
-4	16.8		12.6-.77 -
29.68			
Low-predictability			
+8	23.9	15.9	4.5 -41
+4	26.6	13.2	12 -52
0	21.7	15.6	11.8 -51.79
-4	8.7	12.1	-.77 -29.6

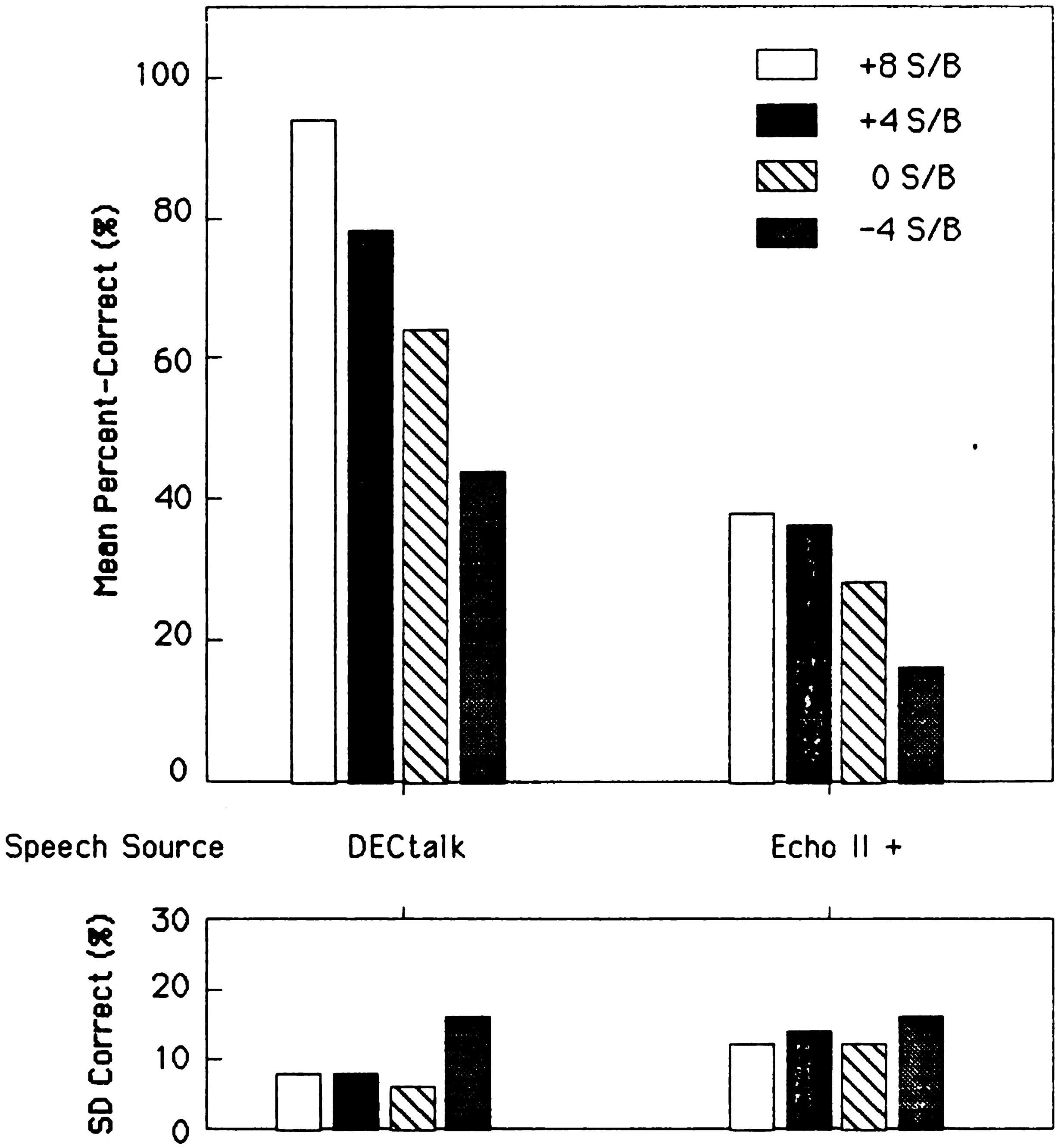


Figure D-1. Means percent-correct and standard deviations for the SPIN full-listsas a function of signal-to-babble ratio using DECtalk and Echo II + as speech sources. Each bar represents observations of 7 normal-hearing subjects tested monaurally. The lower histogram denotes ±1 standard deviation.

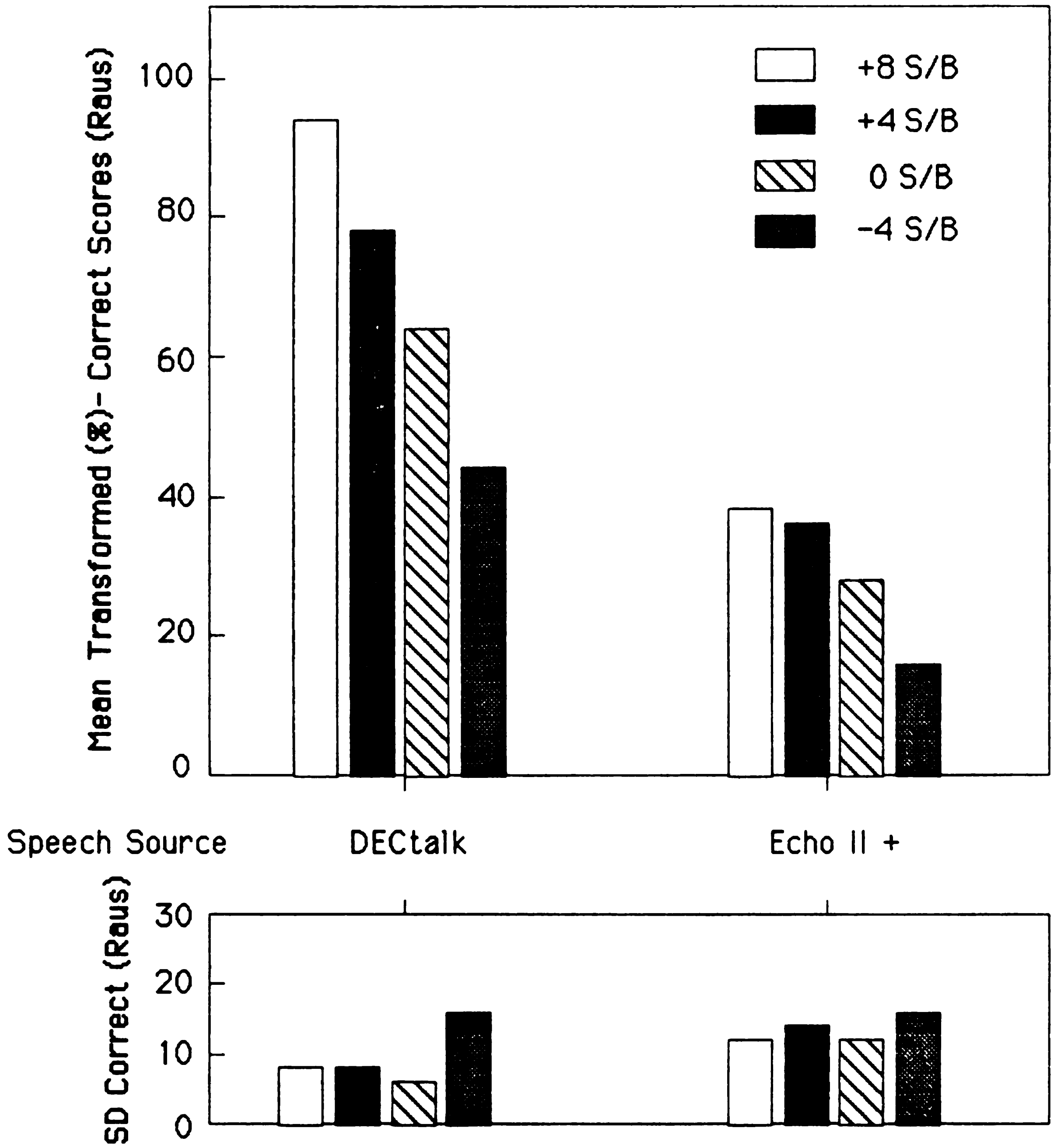


Figure D-2. Mean arcsin transformed percent-correct scores and standard deviations for the SPINas a function of signal-to-babble ratio using Echo II + and DECtalk as speech sources. Each bar represents observations on 7 normal-hearing subjects tested monaurally. The lower histogram denotes ±1 standard deviation.

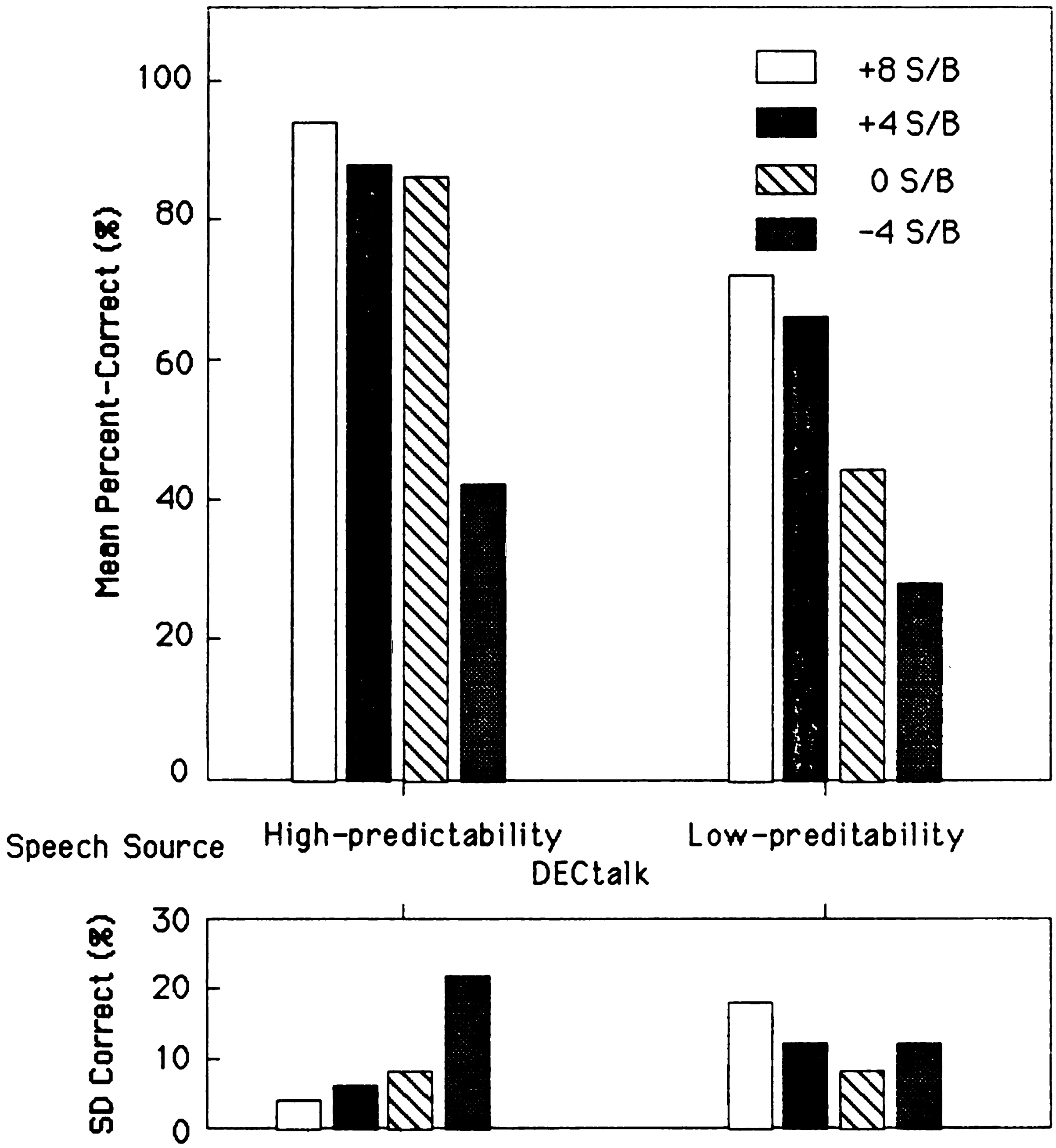


Figure D-3. Mean percent-correct scores and standard deviations for SPIN high and low-predictability test results as a function of signal-to-babble ratio using DECtalk as a speech source. Each bar represents observations of 7 normal-hearing adult subjects. The lower histogram denotes ±1 standard deviation.

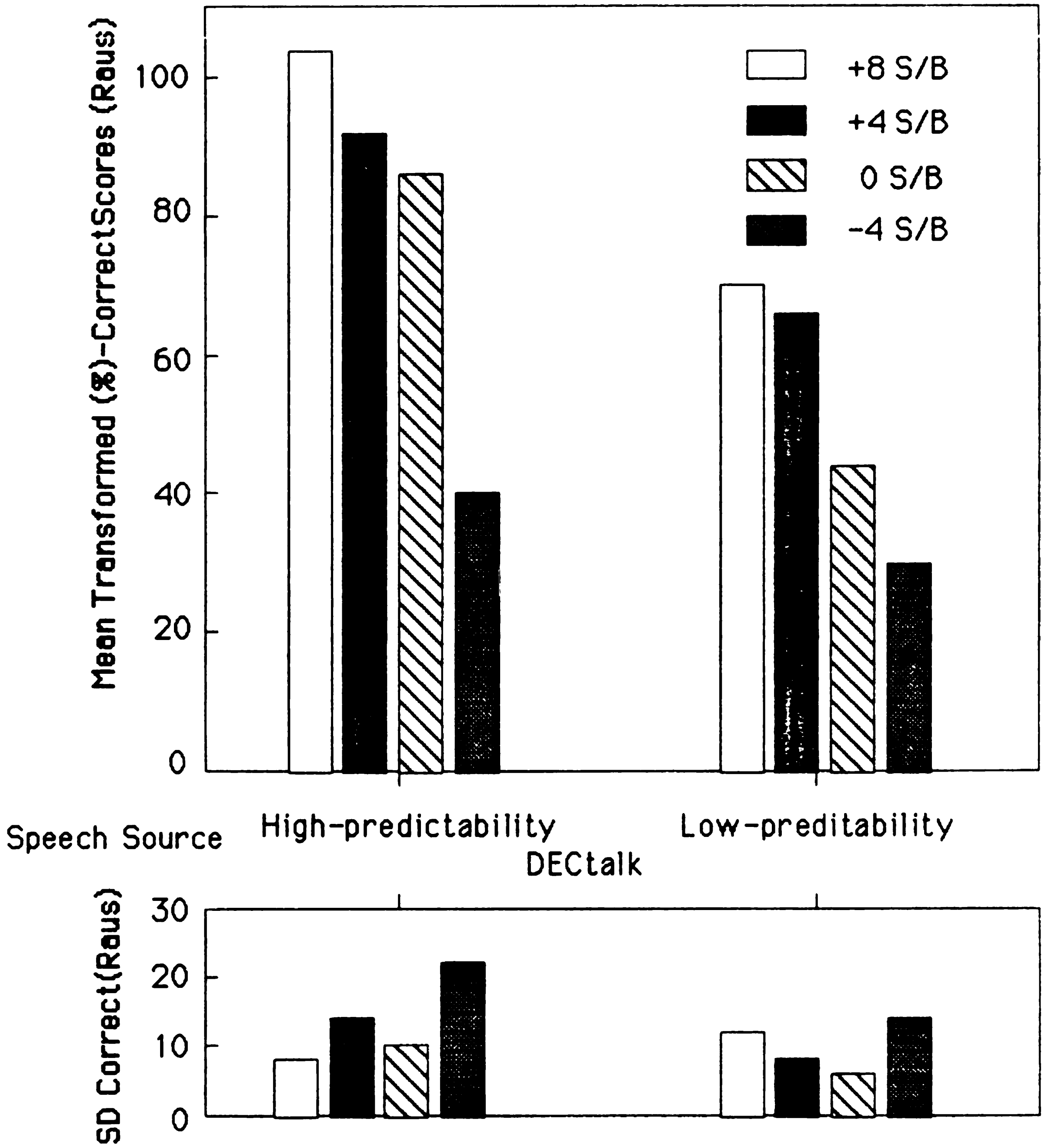


Figure D-4. Mean transformed percent-correct scores and standard deviations for SPIN high and low-predictability test results as a function of signal-to-babble ratio using DECtalk as a speech source. Each bar represents observations of 7 normal-hearing adult subjects. The lower histogram denotes ±1 standard deviation.

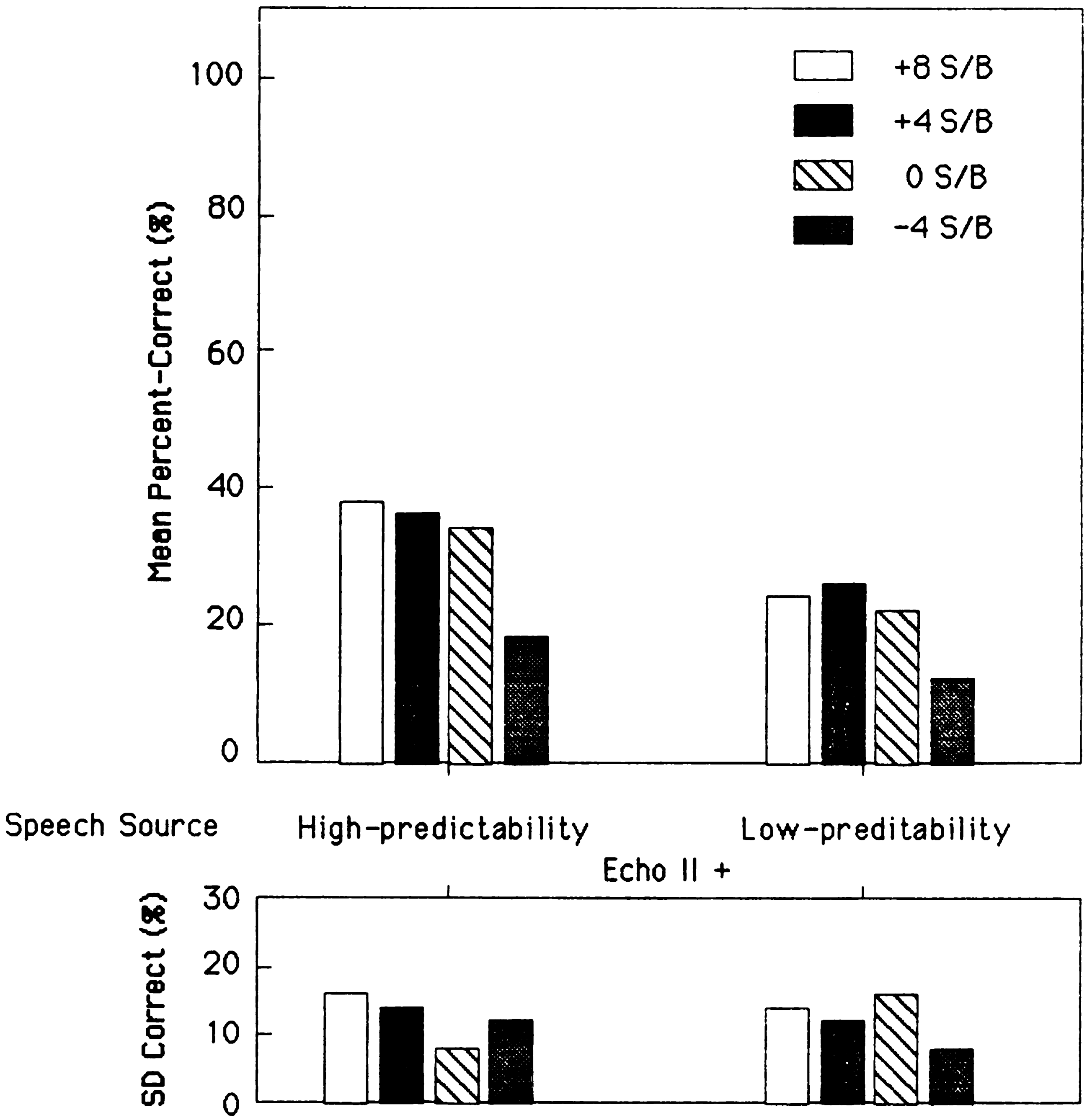


Figure D-5. Mean percent-correct scores and standard deviations for SPIN high and low-predictability test results as a function of signal-to-babble ratio using Echo II + as a speech source. Each bar represents observations of 7 normal-hearing adult subjects. The lower histogram denotes ±1 standard deviation.

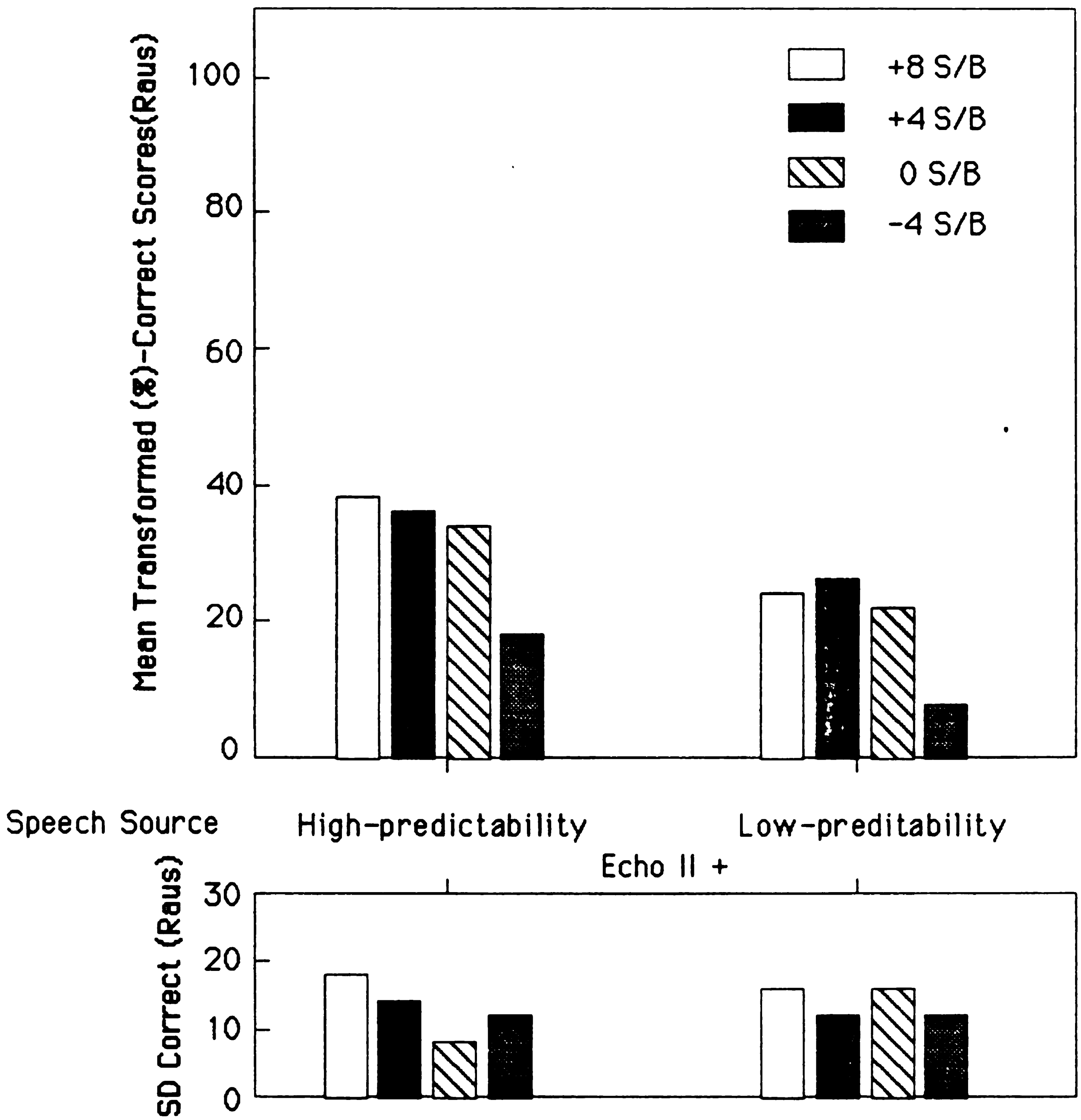


Figure D-6. Mean percent-correct transformed scores and standard deviations for SPIN high and low-predictability test results as a function of signal-to-babble ratio using Echo II + as a speech source. Each bar represents observations of 7 normal-hearing adult subjects. The lower histogram denotes ±1 standard deviation.

APPENDIX E

Appendix E. Summary of Passage Characteristics

Passage One. Lesson Number Ten.

Number of Words	123
Number of Syllables	150
Number of Sentences	9
Mean Words per Sentence	13
Gunning's Fog Index	6
Flesch's Index	94 (grade 6)

Passage Two. Lesson Number Twenty Four.

Number of Words	118
Number of Syllables	158
Number of Sentences	9
Mean Words per Sentence	13
Gunning's Fog Index	7
Flesch's Index	80 (grade 7)

Passage Three. Lesson Number Forty Three.

Number of Words	116
Number of Syllables	146
Number of Sentences	9
Mean Words per Sentence	12
Gunning's Fog Index	6
Flesch's Index	88 (grade 6)

Passage Four. Lesson Number Sixty Eight.

Number of Words	118
Number of Syllables	159
Number of Sentences	9
Mean Words per Sentence	13
Gunning's Fog Index	6
Flesch's Index	81 (grade 6)

Passage Five. Lesson Number Seventy Four.

Number of Words	117
Number of Syllables	146
Number of Sentences	11
Mean Words per Sentence	10
Gunning's Fog Index	6
Flesch's Index	94 (grade 5)

Appendix E. (continued)

Passage Six. Lesson Number Eighty Seven.

Number of Words	134
Number of Syllables	162
Number of Sentences	7
Mean Words per Sentence	12
Gunning's Fog Index	6
Flesch's Index	93 (grade 5)

Practice Passage.

Number of Words	110
Number of Syllables	146
Number of Sentences	7
Mean Words per Sentence	15
Gunning's Fog Index	9
Flesch's Index	79 (grade 7)

Appendix E. (continued)

Table E-1. Data From Connors (p.125, 1974) For The Six Passages Selected for Use in Experiment 3.

Lesson	Grade Level	Listening Score	Difference Score	Ave Sentence Length in Words	Syllabels per 100 Words
10	4.4	.89	.47	13.7	198
24	3.7	.96	.44	14.7	129
43	4.2	.91	.39	12.8	126
68	4.5	.88	.44	13.2	128
74	3.7	.79	.45	10.6	118
87	3.8	.88	.52	12.2	118

Appendix E. (continued)

Practice Passage 1

People in England once imagined that rubies and diamonds lay on the shore of Virginia like pebbles, and the first Indians made their pots and pans of gold. James the First, the English King, wanted hemp for the rigging of his ships and tar for the seams. He was sure Virginia's fields and forests could supply all his needs. Silkworms would grow fat. Grapes would be bigger and sweeter than those of Spain. Figs and olives would thrive in this sunny land.

Fishermen from England saw a different Virginia. They cast their nets in a misty sea and dried them on a rocky coast. Huts and sheds gave them cover from the storms while ashore.

Appendex E. (continued)

Test For Practice Passage 1

Name ID# Order Date

1. King James wanted
   (a) fish for his navy; (b) jewels for his treasury; (c) rigging for his ships; (d) furs for his courtiers.

2. Fishermen from England lived in
   (a) tents; (b) cabins; (c) houses; (d) huts.

3. The English thought Indians made their pots and pans from
   (a) gem stones; (b) gold; (d) clay; (d) wood.

4. The weather in Virginia was thought to be
   (a) sunny; (b) misty; (c) stormy; (d) cold.

5. Grapes would grow bigger than in
   (a) England; (b) Spain; (c) China; (d) France.

6. The English believed many things would grow well in Virginia including;
   (a) silkworms (b) corn (c) wheat (d) cattle.

7. Fisherman saw a Virginia which
   (a) had pleasant beaches; (b)was what they expected; (c) would supply them hemp; (d) had rocky coasts.

8. What did the English believe lay on the shore?
   (a) pebbles; (b) figs; (c) rubies; (d) gold.

Appendix E. (continued)

Practice Passage 2

A door is an opening through which people enter and leave a room or building. The word door also means the movable frame that is used to open and close such an opening. This frame may be hung on hinges. It may slide back and forth in a groove, or it may turn on a pivot like a vertical axle. There are doors that are divided into two parts, so that the upper half can be opened while the lower half stays closed. This is a popular type door in European cottages. The top section is opened to allow the people to see out while the bottom section remains closed.

Appendix E. (continued)

Instructions for Experiment 3 - Listening Comprehension

You are going to hear a recordings of passages made by several different talkers. You will hear the speaker in only one ear. The passages will come from the earphone on your _____ ear. It will be hard to hear them because they will be played in the presence of a background noise of many people talking at the same time. I want you to ignore the noise completely. Pay attention only to the voice.

Once you have listened to the passage you will be given a short test regarding the content. The test will consist of eight or nine multiple choice questions. Do not leave a question blank. If you are not sure of an answer please guess. Keep in mind you will be scored on the time it takes you to complete the test as well as the number correct so work as quickly as you can. I will tell you when to begin and when you are done press this button to stop the timer.

There are 6 passages each lasting approximately 30 seconds. At the end of each multiple choice test there will be a short break. Remember, your job is to listen to the passage and then answer the test questions about the passage as quickly as you can.

Any questions? Let's try a practice set.

Appendix E. (continued)

Passage Number One. Lesson Number Ten.

How do you know when your goldfish are hungry?

A man once taught his goldfish to ring a little bell when they wished for food. He began by letting them go hungry for a few days. Then, before their food was dropped in, it was tied to a string and a bell. As the fish nibbled, the bell outside the bowl rang loudly. For several days they were fed in this way. Then the string without the food was put in and the fish bit at the end just the same. When the bell rang the man threw in some water fleas for the goldfish to eat. After several days of this training they learned to ring the bell whenever they were hungry.

Appendix E. (continued)

Test For Passage Number one, Lesson Number Ten.

Name ID# Order Date

1. The string was tied to the
   (a) food; (b) fish; (c) pebbles; (d) bowl.

2. The bell was
   (a) in the water; (b) on the man; (c) outside the bowl; (d) on the fish.

3. The man fed the fish with
   (a) flies; (b) gold; (c) string; (d) water fleas.

4. The fish were not fed for
   (a) several weeks; (b) a month; (c) a few days; (d) eight days.

5. The bell rang when the fish
   (a) swam; (b) pulled the string; (c) were hungry; (d) drank.

6. The fish were taught to
   (a) swim; (b) nibble food; (c) ring a bell; (d) catch flies.

7. To train the fish required
   (a) several days; (b) two weeks; (c) four weeks; (d) one day.

8. The fish pulled the string when they wished to
   (a) sleep; (b) eat; (c) swim; (d) play.

9. This story tells you how to
   (a) feed fish; (b) catch fish; (c) cook fish; (d) train fish.

Appendix E. (continued)

Passage Number Two. Lesson Number Twenty Four.

Most ants are great fighters and often fight in regular armies. When one army wants to attack an ant hill, it sends scouts ahead and behind, to look for danger. The ants swarm over the ant hill they wish to capture. If they are successful, they carry away the dead bodies of their enemies. They also carry away the eggs of their enemy ants to their own homes. When these eggs are hatched the ants from them become slaves. These slaves work very hard and have little time for rest. Sometimes the ants that are waited on all the time by the slaves become so helpless that they are not able to walk or move, and become invalids.

Appendix E. (continued)

Test For Passage Number Two, Lesson Number Twenty Four

Name ID# Order Date

1. The ants in this story are
   (a) cowards; (b) fighters; (c) drones; (d) friends.

2. Do ants attack
   (a) in pairs; (b) in groups; (c) alone; (d) in armies?

3. When ant attack an ant hill they
   (a) knock it down; (b) swarm over it; (c) surround it; (d) carry it away.

4. When the battle is over what do they carry away?
   (a) sand; (b) food; (c) dead ants; (d) flies.

5. The ant armies look for danger by sending out
   (a) messages; (b) scouts; (c) alarms; (d) signals.

6. What do the ants do with the eggs of their enemies?
   (a) destroy them; (b) eat them; (c) carry them home; (d) bury them.

7. The ants that are hatched from the enemies' eggs become
   (a) leaders; (b) fighters; (c) friends; (d) slaves.

8. Ants that are always waited on by others often become
   (a) selfish; (b) lazy; (c) helpless; (d) busy.

Appendix E. (continued)

Passage Number Three. Lesson Number Forty Three.

Did you ever walk across a lawn and find a little ridge where there had been none before? Did you wonder where it came from? It was probably made by a little animal about six inches long called a mole. Few people ever see him, for he always stays under ground making tunnels. He has thick, soft gray fur and a short tail which looks like an angle worm. His eyes are so tiny that you can find them only by parting the fur around them. They were made to see only light and dark. He does not need them to see anything else. If they were larger they would always be getting full of dirt.

Appendix E. (continued)

Test For Number Passage Three Lesson Number Forty Three

Name ID# Order Date

1. What does the mole make on the lawn?
   (a) hole; (b) ridge; (d) ditch; (d) hill.

2. He always stays
   (a) in a tree; (b) in a hollow log; (c) in the pond; (d) under the ground.

3. His fur is
   (a) long and black; (b) course and brown; (c) soft and gray; (d) white and shaggy.

4. How long is the mole?
   (a) ten inches; (b) six feet; (c) six inches; (d) two feet.

5. His eyes are
   (a) bright; (b) full of dirt; (c) large; (d) small.

6. The mole is seen by
   (a) few people; (b) many beople; (c) everybody; (d) nobody.

7. He can see
   (a) very well; (b) just a little; (c) everything around him; d) nothing.

8. His tail looks like
   (a) an angleworm; (b) a brush; (c) a shovel; (d) a snake.

Appendix E. (continued)

Passage Number Four. Lesson Number Sixty Eight.

Do you ever wonder where the sugar in your sugar bowl comes from? It may come from one of two plants, the sugar beet or the sugar cane. The sugar cane grows only in warm countries, but the sugar beet grows in states as far north as Michigan.

Sugar cane, when it is growing, looks very much like corn. It is tall and has a jointed stalk. The leaves are long and blade-like. Sugar beets look something like large white turnips. About half of the sugar we use comes from sugar cane and the other half from sugar beets. Some people think cane sugar is better than beet sugar, but there is really very little difference between them.

Appendix E. (continued)

Test For Passage Number Four Lesson Number Sixty Eight

1. The sugar in your sugar bowl comes from
(a) animals; (b) plants; (c) trees; (d) bushes.

2. Sugar cane grows in
(a) Michigan; (b) cold climates (c) warm climates;
(d) almost every country.

3. Sugar beets can grow
(a) only in the south; (b) only were sugar cane grows; (c) in Michigan;
(d) in Maine.

4. When sugar cane is growing, it looks like
(a) corn; (b) turnips; (c) beets; (d) wheat.

5. Sugar beets look like
(a) wheat; (b) corn; (c) potatoes; (d) turnips.

6. What kind of leaves has sugar cane?
(a) long; (b) short; (c) round; (d) curly.

7. How much of our sugar comes from beets?
(a) 1/4; (b) 3/4; (c) 1/2; (d) 1/3.

8. Some people think that the sugar form beets is
(a) better than cane sugar; (b) as good as the sugar form cane;
(c) more expensive than sugar cane; (d) not as good as cane sugar.

9. Between cane and beet sugar, there is really
(a) little difference; (b) much difference;
(c) no difference; (d) a great difference in price.

Appendix E. (continued)

Passage Number Five. Lesson Number Seventy Four.

Most birds make nests, but the nests are not all alike. Every bird has its own kind of nest. The tailor bird sews leaves together for its nest. It sews them with thread which it makes of cotton from the cotton plant. It uses its bill for a needle.

The robin builds its nest of many things. It makes a framework of twigs and sticks, and then plasters it with mud. When this is done it lines the inside of the nest with fine moss, feathers, and hair.

Barn swallows build their nests of mud. They make them in barns, close to the roof. Sometimes several swallows build their nests in a row quite near one another.

Appendix E. (continued)

Test For Passage Number Five Lesson Number Seventy Four

Name ID# Order Date

1. These swallows build nests in
   (a) trees; (b) bushes; (c) grass; (d) barns.

2. The outside of the robin's nest is made of
   (a) leaves; (b) hair and moss; (c) twigs and sticks; (d) feathers.

3. A bird that plasters this nest is the
   (a) bluebird; (b) robin; (c) tailor bird; (d) blue jay.

4. In building its nest the tailor bird uses
   (a) mud; (b) sticks; (c) twigs; (d) leaves.

5. For a needle the tailor bird uses
   (a) his bill; (b) a small twig; (c) his feet; (d) his tongue.

6. Birds that build near each other are
   (a) robins; (b) bluebirds; (c) barn swallows; (d) tailor birds.

7. Tailor birds get thread from
   (a) cloth; (b) a factory; (c) string; (d) the cotton plant.

8. Which would be the best title for this story?
   (a) Birds; (b) Bird's Nests; (c) Kinds of Nests; (d) Nests.

Appendix E. (continued)

Passage Number Six. Lesson Number Eighty Seven.

My father showed me a cracked blue bead and told me this story about it.

I was on a visit in the state of Washington. One day in the early morning I started to walk up a mountain. A new road was being built and a gang of workmen was already at work with shovels and picks. As I came near them, one of the men gave a shout and stooped down to the ground. I, with the rest, hurried to see what he had found. We saw that his pick had opened an old Indian grave. At the first touch, however, most of the bones crumbled into dust. Partly hidden in the dirt were two beads of bright blue. The first man picked them up and gave one to me. This is that bead.

Appendix E. (continued)

Test For Passage Number Six. Lesson Number Eighty Seven

Name ID# Order Date

1. The walk was taken
   (a) in the afternoon; (b) at noon; (c) at night; (d) in the morning.

2. The beads were
   (a) new; (b) blue; (c) round; (d) heavy.

3. The man found
   (a) shovels; (b) a big stone; (c) a grave; (d) a pick.

4. The workmen were
   (a) digging; (b) shouting; (c) stooping; (d) walking.

5. My father father told me a story about a
   (a) mountain; (b) state; (c) road; (d) bead.

6. The beads were covered by
   (a) dirt; (b) leaves; (c) stones; (d) grass.

7. The workman kept
   (a) two beads; (b) one bead; (c) both beads; (d) none.

8. Those who ran to see the discovery were
   (a) everyone; (b) the gang; (c) the foreman; (d) the visitor.

9. Who was on a visit?
   (a) myself; (b) a workman; (c) my father; (d) an Indian.

APPENDIX F

Appendix F.

Multiple Instructions Test. Practice Form.

Number one. Place an X in the yellow rectangle, connect the oval to the second shape and to the red square and circle the triangle.

Number three. Put the Letter K in the diamond and N beneath the circle.

Number twenty two. Place the number three inside the big circle, connect the first and second circle with a line touching the top of the square, cross out the oval and the diamond.

Number twenty four. Write a V left of the cow, a line around the house and underline the girl.

Number two. Underline the boy, write the letter A left of the deer, the number five under the house and above the car.

Number six. Put a square around the R, a circle below the V and connect the S with the Y.

Number thirteen. Circle the letters NP and draw a line from AB to NP.

Number fifteen. Draw a line from the blue to the yellow shape, the letter Q touching the top of the diamond, an A in the rectangle and a line through the oval.

Number four. Circle the word fist and connect the word cat to the shape.

Number nine. Underline the J, draw a square around the L and connect the C to the M.

Number sixteen. Underline the pie, put an M beneath the truck, put an X beside the fish's tail and circle the first star.

Number twenty three. Put a square around the last circle, underline the first, connect the words draw and pen with a line going through top and put a D below the purple.

Number five. Draw a line from the four to the second number that does not touch any other shape, underline the diamond, put an X in the circle and above the square.

Number ten. Put a five in the rectangle and cross out the circle.

Appendix F. (continued)

Multiple Instructions Test. Practice Form. (continued)

Number nineteen. Underline the spoon, draw a circle after the hat and write the letter H over the glass.

Number twenty one. Draw a square left of the diamond, a line over the blue shape, put a P on the oval and under the small rectangle.

Number seven. Write an F in the blue shape that touches the sides, a four to the right side of the green diamond, cross out the oval and circle the square.

Number eleven. Place a two below the girl, circle the animal and connect the word cat with the car.

Number seventeen. Cross out the glass and put a square around the apple.

Number twenty. Circle the fourth letter, cross out the yellow triangle, connect the diamond and red square and the numbers one and eight.

Number eight. Draw a line between the first and fifth shapes, a one under the tree, underline the eight and cross out the apple.

Number twelve. Connect the square to the diamond, put a G inside the rectangle and a six beneath the circle.

Number fourteen. Connect the first and third circles by going through the second, cross out the green shape, write a ten under the yellow circle and a seven under the red circle.

Number eighteen. Circle the five and connect it to the bottom of thirty six.

Appendix F. (continued)

Multiple Instructions Test. Form Number Two.

Number one. Draw a square inside the oval, draw a line above the deer, put an A between the house and spoon, cross out the tree, connect the apple and glass.

Number two. Draw a nine inside the small oval, cross out the third shape, draw a diamond in the rectangle that does not touch the side.

Number three. Place a square above the star, cross out the truck, draw a line under the car, place a two beneath the first object, draw a circle around the fish.

Number four. Draw a G inside the triangle, place a circle under the rectangle, put a line through the diamond.

Number five. Cross out the circle, draw a square around the small circle, place an N in the blue square, underline the yellow square.

Number six. Circle the hat and cross out the number ten.

Number seven. Draw a Z under the first square, underline the rectangle, cross out the last diamond, put a star on the second square.

Number eight. Put an S on the cow and circle the girl.

Number nine. Connect the triangle and circle, cross out the star.

Number ten. Cross out the orange square, underline the sixth square, connect the yellow and green squares, circle the word free, place an H between the words free and top.

Number eleven. Place the number ten in the triangle, cross out the house, draw a line under the cow.

Number twelve. Connect the triangle and star, circle the diamond, draw a four between the triangle and square.

Number thirteen. Underline the truck, cross out the apple, draw a line connecting the tree and the glass that must go below the objects, draw a star on top of the tree.

Number fourteen. Circle the pie, connect the square and the rectangle, draw a three under the house.

Appendix F. (continued)

Multiple Instructions Test. Form Number Two (continued)

Number fifteen. Cross out the small star, draw a square around the knife, put a square inside the rectangle, connect the biggest star to the car.

Number sixteen. Draw a diamond above the green circle, put a rectangle over the third circle, connect jail to fist with a line above the words, put a zero under the fourth word.

Number seventeen. Draw a line between the orange and black shapes going above the shapes, put a star between the square and circle, circle the white rectangle, underline the diamond.

Number eighteen. Place a U in the circle, underline the oval, put a triangle after the square.

Number nineteen. Circle the tree, draw a line between the knife and pie, put a star in the square.

Number twenty. Cross out the rectangle, write the letter B left of the triangle.

Number twenty one. Draw a line between the small star and the rectangle, connect the stars, draw a circle around the square, underline the fourth shape.

Number twenty two. Put a J between the fish and pie, circle the eight.

Number twenty three. Connect the first large and the second small star, underline the square, circle the last two small stars, place a three in the square.

Number twenty four. Draw a line between the truck and boy, circle the E.

Appendix F. (continued)

Multiple Instuctions Test. Form Number Three.

Number one. Connect the second and third circles, put a square around the word cat, an H on open and a line under the number two.

Number two. Write a four between the pink and yellow shapes, a line through the triangle that touches the circle, an R to the right of the triangle and underline the square.

Number three. Place the number three beside the girl and cross out the fork.

Number four. Circle the letters inside the second rectangle, connect the first and second circle, put an X above number fourteen and B below thirty six.

Number five. Connect the star with the small rectangle, draw a square above the red shape, circle the oval and underline the blue shape.

Number six. Write a D beside the ball, a line around the spoon and cross out the man.

Number seven. Cross out the oval and connect the diamond to the triangle.

Number eight. Write a C in the large shape, an F below the pie, underline the diamond and draw a square on top of the cow.

Number nine. Connect the large and small triangles with a line that goes above the shapes, put an X between the purple and green shapes, underline the diamond and the oval.

Number ten. Make a square beside the rectangle and cross out the letter R.

Number eleven. Circle the fork and place a square in front of the fish.

Number twelve. Put a two to the right of the car, connect the triangle to the pie, circle the star and underline the seven.

Number thirteen. Draw a line under the tree, cross out the house, put an N in the diamond.

Number fourteen. Place an X below the circle, draw a line from the orange diamond to the square that touches the blue diamond, put an eight in the rectangle and the triangle.

Appendix F. (continued)

Multiple Instructins Test. Form Number Three (continued)

Number fifteen. Connect the pink and black squares, circle the truck, underline the spoon and write the letter Q to the left of the diamond.

Number sixteen. Make a square around the small circle that touches the sides, connect the white square and the diamond, put a V in the oval and a five in the circle.

Number seventeen. Draw a line between the two words, cross out the hat and put a square under the deer.

Number eighteen. Connect the fish and car, put a line under the rectangle.

Number nineteen. Cross out the knife and connect the tree to the apple.

Number twenty. Put an O in the small square and cross out the two large squares.

Number twenty one. Draw a square around the star, an X in the smallest shape that touches the side, connect the purple and green shapes and put an E in the rectangle.

Number twenty two. Draw a line between the tops of the blue and yellow shapes, a diamond in the square, underline the first triangle and the circle, connect the circle and second triangle .

Number twenty three. Make a circle inside the diamond, connect the large and small rectangles and cross out the star.

Number twenty four. Make a K over the car, a P left of the apple and underline the tree.

Appendix F. (continued)

Mulitple Instructions Test Test. Form Number Four.

Number one. Cross out the biggest triangle, write an R on top of the smallest, put an X beside the white shape and inside the yellow.

Number two. Draw a line from the purple shape to the star that goes around the oval, put a nine left of the triangle, underline the square and green shape.

Number three. Cross out the apple and put a square next to the knife.

Number four. Make a triangle next to the star, circle the car, connect it to the bottom of the pie and underline the number three.

Number five. Connect the first and fourth square with a line that goes through number ninety, write a K in the last square, cross out the NP and underline the twenty three.

Number six. Write an O above the house and cross out the cow.

Number seven. Draw a square around the triangle, circle the P and connect the number eight and the diamond.

Number eight. Draw a line from the purple shape to the D that goes under the C, underline the man, put an X under the F, and another before the truck.

Number nine. Make a circle around the fourteen and a G on the diamond.

Number ten. Put a circle in the triangle that touches the sides, draw a line from the pink to the orange shape that goes through the black and cross out the oval.

Number eleven. Connect the top of the house to the car, the cow to the deer and circle the man.

Number twelve. Write an N next to the glass, place a line beneath the tree, an X over the hat and connect the truck and apple.

Number thirteen. Draw a line from the star to the ball that circles the pie and cross out the tree.

Number fourteen. Make a square around the fish, a five in the rectangle and an X on the fork.

Appendix F. (continued)

Multiple Instructions Test. Form Number Four. (continued)

Number fifteen. Underline the small circle, connect the blue square with the second circle using a line over the shapes, write two in the small square and D on the green shape.

Number sixteen. Cross out the oval and write a J in the rectangle.

Number seventeen. Circle the big star and connect it to the top of the knife, underline the diamond.

Number eighteen. Draw a line from the small white shape to the star that goes through the third shape, connect the large and small rectangles, underline the star and cross out the oval.

Number nineteen. Write an L below the diamond and a three in the rectangle.

Number twenty. Draw a square around pen, a circle around top, put an X left of the yellow circle and write a Q in the third circle.

Number twenty one. Connect the spoon to the house, underline the car and place a two beneath the triangle.

Number twenty two. Make a square around the black shape, a circle around the pink shape, cross out the blue shape and put a Z in the oval.

Number twenty three. Write a two left of the square and underline the triangle.

Number twenty four. Put an X on the glass, a P on the blue shape, a line under the spoon and a one right of the diamond.

Appendix F. (continued)

Multiple Instructions Test Test. Form Number Five.

Number one. Connect the small square and the glass, underline the car, put a five on the right of the fork, cross out the large shape.

Number two. Draw a line between the yellow shape and the first small triangle, circle the circle, put an X under the rectangle, underline the purple square, cross out the circle.

Number three. Draw a square around the two, connect the ten and six.

Number four. Place an I in the green diamond, connect the rectangle and the yellow shape, draw an oval between the second diamond and the square, circle the rectangle.

Number five. Circle the fish, draw a Q between the pie and the house, cross out the cow's tail.

Number six. Draw a star to the left of the girl, connect the deer and car, put a circle around the boy, underline the cow.

Number seven. Make a V below the oval, connect the star and small rectangle, cross out the large rectangle.

Number eight. Circle the spoon, draw a line between the car and the spoon.

Number nine. Underline the first diamond, connect the blue and pink shapes, put a B in the white diamond, cross out the circle.

Number ten. Draw a line between the orange and blue shapes that goes through the yellow circle, underline the green shape, cross out the white and middle circles.

Number eleven. Connect the first and last circle, draw an L under and a seven above cat, put a star over door, underline jail.

Number twelve. Cross out the star, put a zero in the rectangle.

Number thirteen. Connect the large circle and the small square, underline the oval, cross out the diamond.

Number fourteen. Draw a circle around the six, underline the glass, connect the fish's tail and the hat.

Appendix F. (continued)

Multiple Instructions Test. Form Number Five. (continued)

Number fifteen. Draw a line between the purple shapes, put an X under the first and last shapes, make a circle between the square and diamond.

Number sixteen. Circle the apple, draw a C to the right of the deer.

Number seventeen. Underline the smallest shape, draw a star under the orange shape, cross out the oval, make an F in the rectangle.

Number eighteen. Put a ten above the small circle, underline the last shape, connect the yellow and orange shapes, draw a diamond around the first shape, cross out the square.

Number nineteen. Connect the ball and the spoon, put a square under the car.

Number twenty. Cross out the square and make an R above number twenty three.

Number twenty one. Connect the first and last words, make a W above the girl, cross out the car, make an oval to the left of the house, circle the second object.

Number twenty two. Put a J behind the fish, underline the rectangle, draw a line between the star and triangle.

Number twenty three. Connect the small star and diamond, underline the orange shape, cross out the yellow shape, put a C above and below the rectangle.

Number twenty four. Place a K on the truck, underline the fork, make the number three below the hat.

Appendix F. (continued)

Multiple Instructions Test Test. Form Number Six.

Number one. Put a D in the triangle, a number four in the diamond and underline the rectangle.

Number two. Draw a line from door to bus, underline cat and put a U in the shape.

Number three. Underline the first word, connect the ball to the second word, put a line beside the hat and a G over the deer.

Number four. Make a square around the house, a circle around the boy, put a nine below the car and connect the third and fourth objects.

Number five. Put an X under the fish's tail and cross out the pie.

Number six. Connect the two black shapes with a line that circles the square, write an M in the rectangle and underline the triangle.

Number seven. Draw a line through the Y, connect the N to the S and cross out the U.

Number eight. Circle the oval and write a two in the triangle.

Number nine. Make a square around the third big star and connect it to the second small star, circle the number, underline the square and cross out the first star.

Number ten. Write a four in the square , an X over the girl and put a two below the apple.

Number eleven. Put a square around the yellow shape that does not touch it, write a Z left of the second shape, a five above the green shape and underline the diamond.

Number twelve. Connect the orange shape to the white rectangle, write an R between the hat and the fork, place a square behind the fish's tail, connect all the shapes.

Number thirteen. Draw a line beside letter A and put a line through E.

Number fourteen. Write an eight in the second triangle, a four in the orange shape, connect the yellow triangle with the square with a line over the shapes and underline the circle.

Appendix F. (continued)

Multiple Instructions Test. Form Number Six.

Number fifteen. Make a triangle around the deer, a line through the fork and an X right of the apple.

Number sixteen. Circle number thirty six and draw a square around number one.

Number seventeen. Write a J in the oval, a circle over the rectangle and a line through the triangle.

Number eighteen. Connect open to can with a line that circles the oval, cross out the two, draw a square around the second circle and a diamond to the left of cat.

Number nineteen. Put a B in the green shape, a five in the small white shape, circle the triangle and cross out the orange rectangle.

Number twenty. Draw a line from the blue square to the yellow diamond that does not touch another shape, cross out the biggest rectangle, circle the oval and underline the star.

Number twenty one. Connect the spoon to the apple and circle the glass.

Number twenty two. Cross out the knife and put an M under the tree.

Number twenty three. Put an X in the orange shape and an E in the diamond, circle the square and connect the two ovals with a line that goes through the rectangle.

Number twenty four. Circle the purple triangle, draw a square between the green and blue shapes and put an X over the star and inside the rectangle.

Appendix F. (continued)

Multiple Instructions Test. Form Number Eight

Number one. Make a line from the two to the six, connect the second rectangle to the knife, underline the two and circle the hat.

Number two. Draw a square under the purple shape, a G beside the oval, a line between the orange and yellow shapes and around the star.

Number three. Circle the AB, connect it to the NP and draw a line between thirty six and fifty seven.

Number four. Place a five in the circle, a three below the diamond, a line between the pink and the green shapes that goes through the blue and cross out the rectangle.

Number five. Put an F left of the apple, a T on the glass, a circle around the woman and underline the tree.

Number six. Put a seven between the stars and cross out the car.

Number seven. Cross out the smallest shape, connect the pink and the yellow shapes without touching any others, put an L in front of the cow and a four over the house.

Number eight. Connect the third object with the fourth object, write an eight beneath the back of the deer and draw a line through the ball.

Number nine. Place an X on the pie and circle the glass.

Number ten. Write a six in the rectangle, a B in the diamond and a ten in the triangle.

Number eleven. Put a line over the number nine, a circle around the number fourteen, connect the third and fifth numbers and cross out number fifty seven.

Number twelve. Connect the black square to the second small triangle with a line going around the purple triangle, underline the small circle, put a Q in the circle.

Number thirteen. Make a rectangle around the pie, a line from the star to the seven going over the ball, a J behind the car and a P under the triangle.

Number fourteen. Cross out the fork, write a four before the truck and an X next to the girl.

Appendix F. (continued)

Multiple Instructions Test. Form Number Eight. (continued)

Number fifteen. Circle the word free and connect it to the third square, put a C in the second, a nine in the purple square and connect the words bus and top.

Number sixteen. Put a line over the circle and cross out the triangle.

Number seventeen. Place an X on the ball, an eight in front of the deer and underline the apple.

Number eighteen. Draw a line from the first shape to the yellow going under the shapes, a circle under the small rectangle, a star under the purple and a D in the oval.

Number nineteen. Circle the F and connect the man to the truck.

Number twenty. Make a circle around the diamond and cross out the oval.

Number twenty one. Write a two in the square, an L in the rectangle and a line over the star.

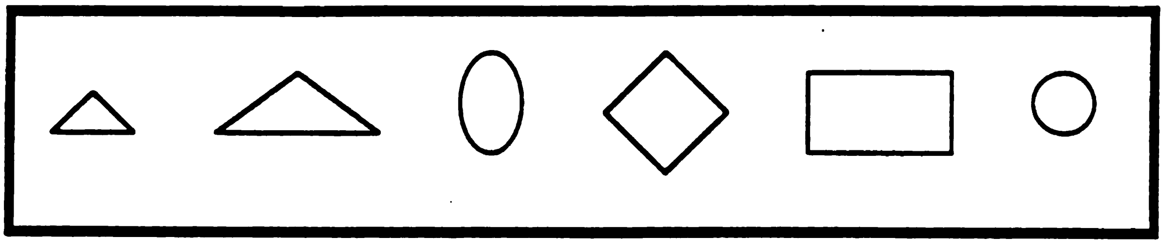
Number twenty two. Connect the triangle to the square and the oval to the diamond.

Number twenty three. Put a line between the blue and the orange shapes, a line through the two, underline can and cross out the second number.

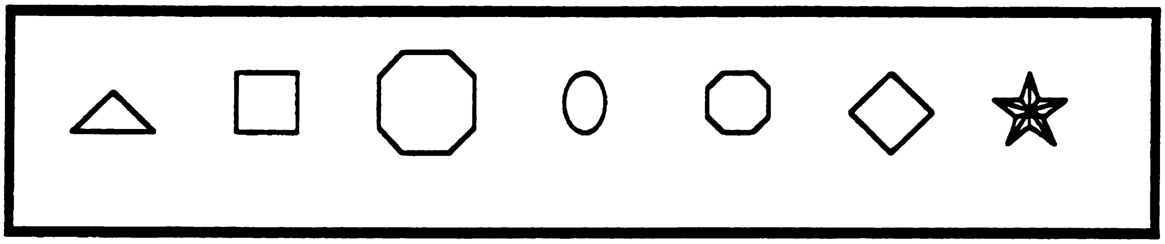
Number twenty four. Make a square around the house, an X beside the ball and a circle beneath the apple, circle the five and connect it to the bottom of thirty six.

Appendix F. (continued) Sample of MIT response form number four.

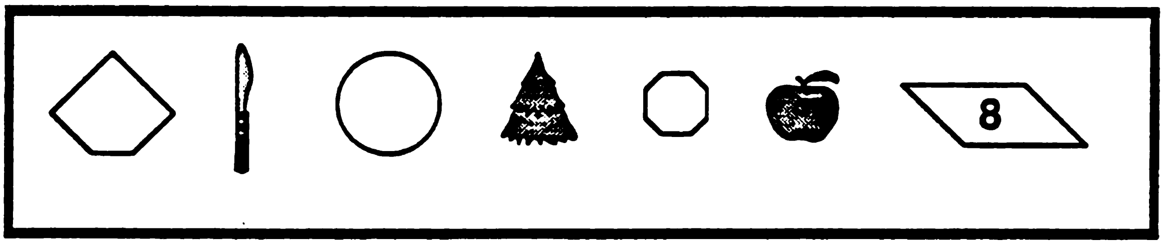
1.



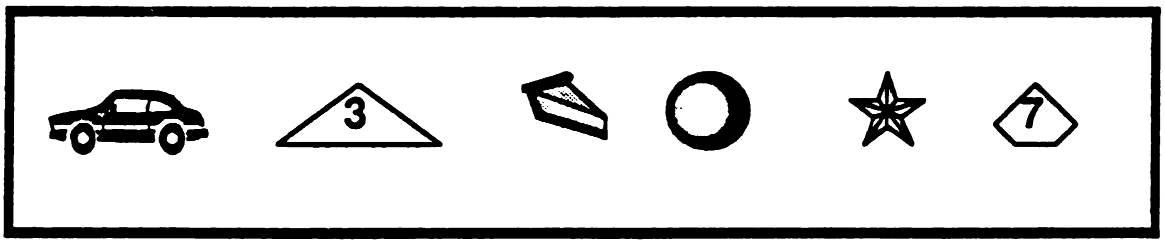
2.



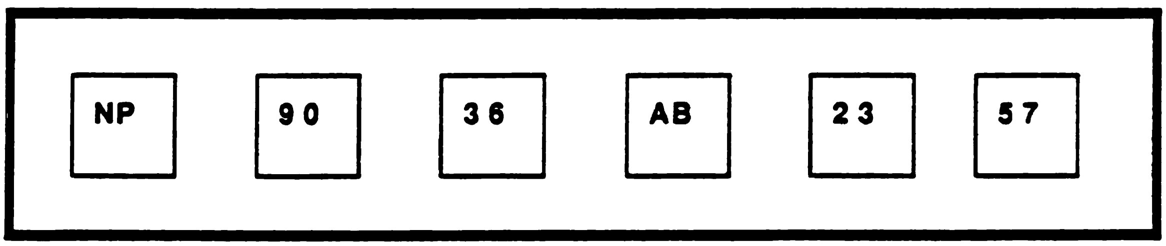
3.



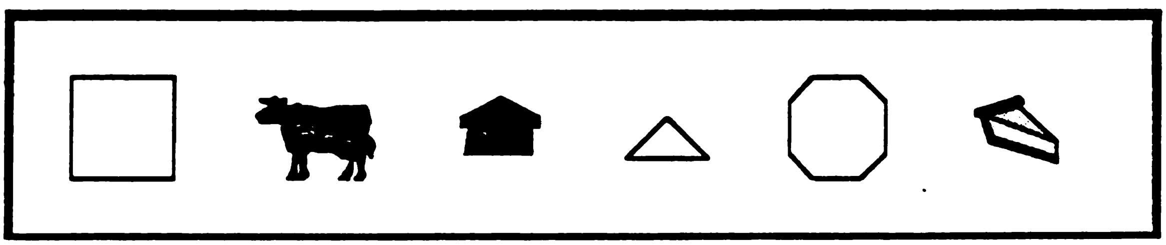
4.



5.

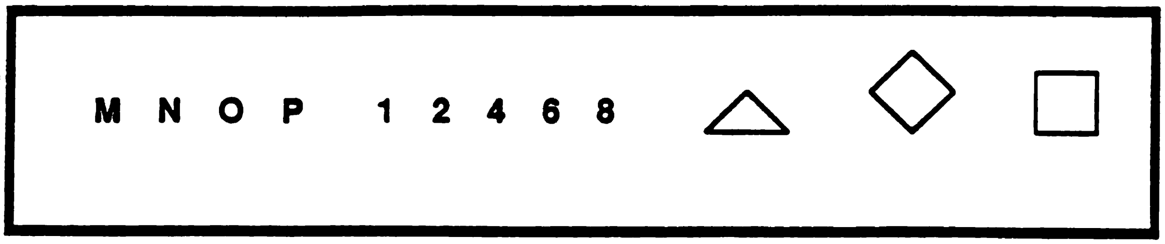


6.

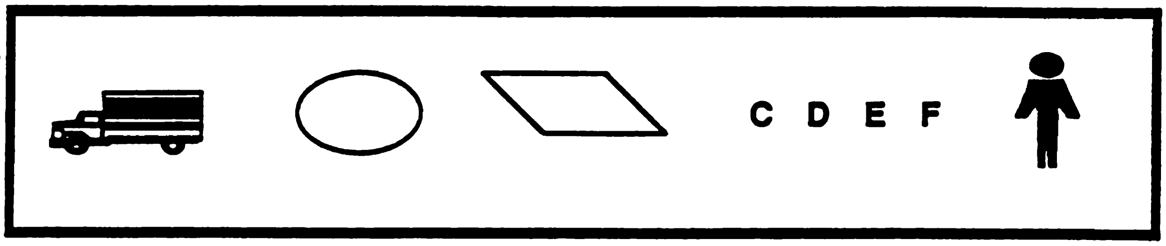


Appendix F. (continued) Sample of MIT response form four.

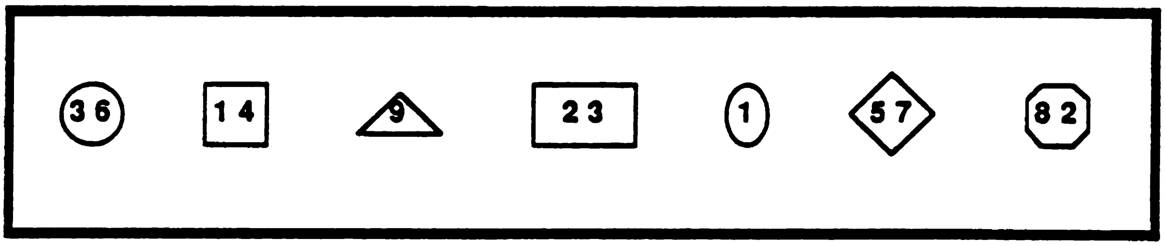
7.



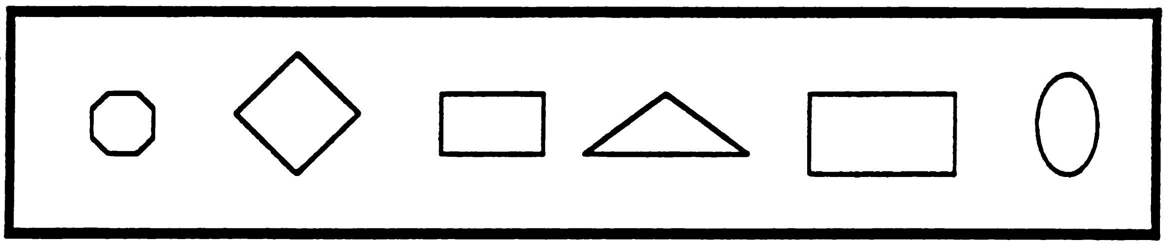
8.



9.

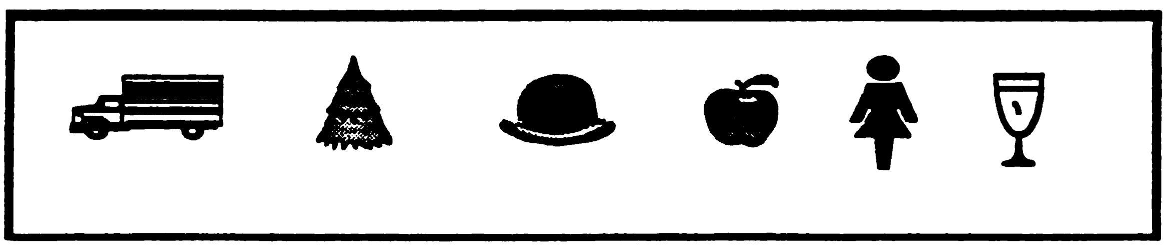


10.



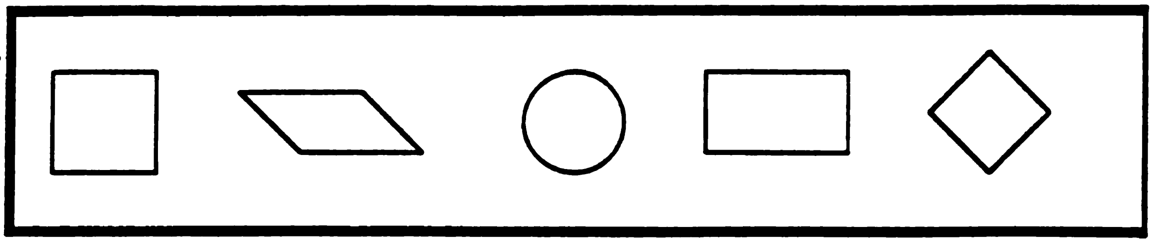
11.

12.

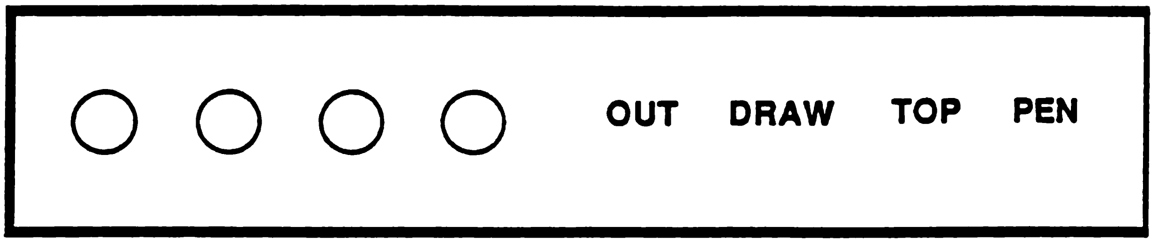


Appendix F. (continued) Sample of MIT response form number four.

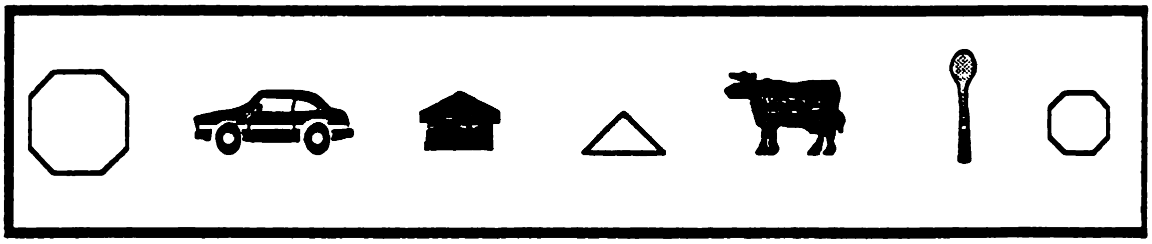
19.



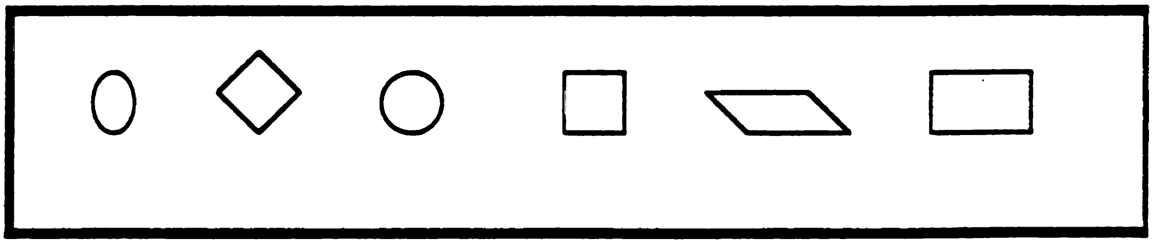
20.



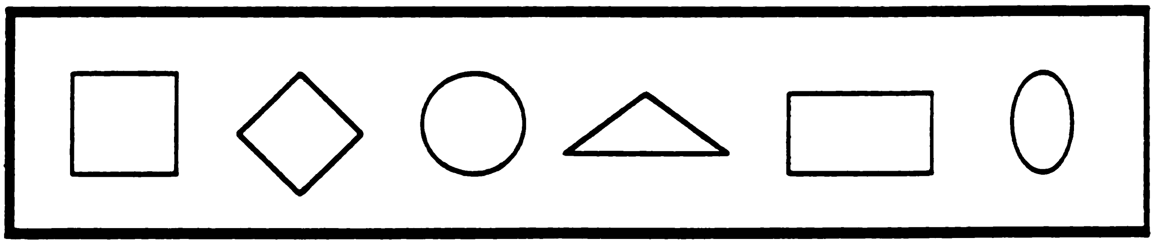
21.



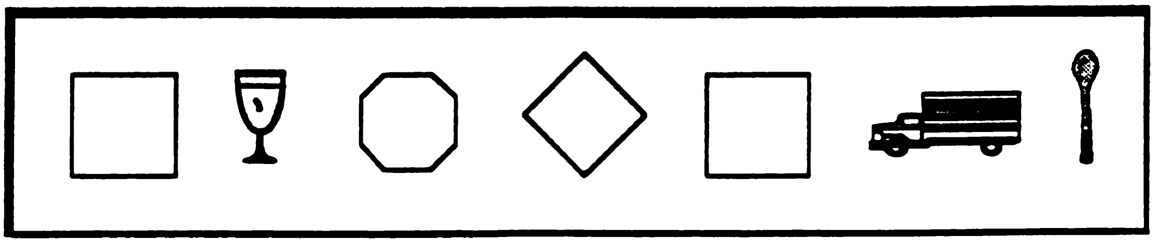
22.



23.



24.



Appendix F. (continued)

Multiple Instructions Test. Score Form number one.

			Total
1.	Place an X in the yellow rectangle.	_____	
	Connect the oval to the second shape.	_____	
	Oval to the red square.	_____	
	Circle the triangle.	_____	_____
2.	Underline the boy.	_____	
	Write the letter A left of the deer.	_____	
	Number five under the house.	_____	
	Above the car.	_____	_____
3.	Letter K in the diamond.	_____	
	N beneath the circle.	_____	_____
4.	Circle the word fist.	_____	
	Connect the word cat to the shape.	_____	_____
5.	A line from the four to the 2nd number.	_____	
	Does not touch any other shape.	_____	
	Underline the diamond.	_____	
	X in the circle.	_____	
	Above the square.	_____	_____
6.	Put a square around the R.	_____	
	Circle below the V.	_____	
	Connect the S with the Y.	_____	_____
7.	Write an F in the blue shape.	_____	
	Touches the sides.	_____	
	Four to the right side of the green diamond.	_____	
	Cross out the oval.	_____	
	Circle the square.	_____	_____
8.	Draw a line between the first and fifth shapes.	_____	
	A one under the tree.	_____	
	Underline the eight.	_____	
	Cross out the apple	_____	_____
9.	Underline the J.	_____	
	Draw a square around the L.	_____	
	Connect the C to the M.	_____	
10.	Put a five in the rectangle.	_____	
	Cross out the circle.	_____	_____
11.	Place a two below the girl.	_____	
	Circle the animal.	_____	
	Connect the word cat with the car.	_____	_____

12.	Connect the square to the diamond.	______	
	G inside the rectangle.	______	
	Six beneath the circle.	______	______
13.	Circle the letters NP.	______	
	Line from AB to NP.	______	______
14.	Connect the first and third circles.	______	
	By going through the second.	______	
	Cross out the green shape.	______	
	Ten under the yellow circle.	______	
	Seven under the red circle.	______	______
15.	Line from the blue to the yellow shape.	______	
	Letter Q	______	
	Must touch the top of the diamond.	______	
	A in the rectangle.	______	
	Line through the oval.	______	______
16.	Underline the pie.	______	
	M beneath the truck.	______	
	X beside the fish's tail.	______	
	Circle the first star.	______	______
17.	Cross out the glass.	______	
	A square around the apple.	______	______
18.	Circle the five.	______	
	Connect it to the bottom of thirty six.	______	______
19.	Underline the spoon.	______	
	Circle after the hat.	______	
	H over the glass.	______	______
20.	Circle the fourth letter.	______	
	Cross out the yellow triangle.	______	
	Connect the diamond and red square.	______	
	Numbers one and eight.	______	______
21.	Square left of the diamond.	______	
	Line over the blue shape.	______	
	P on the oval.	______	
	P under the small rectangle.	______	______
22.	Three inside the big circle.	______	
	Connect the first and second circle.	______	
	With a line touching the top of the square.	______	
	Cross out the oval.	______	
	Cross out diamond.	______	______

23. Square around the last circle. _____
Underline the first. _____
Connect the words draw and pen. _____
With a line going through top. _____
D below the purple. _____ _____

24. V left of the cow. _____
Line around the house. _____
Underline the girl. _____ _____

Total/X/SD _____ _____ _____

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Two.

			Total
1.	Draw a square inside the oval.	______	
	Line above the deer.	______	
	A between the house and spoon.	______	
	Cross out the tree.	______	
	Connect the apple and glass.	______	______
2.	Draw a nine inside the small oval.	______	
	Cross out the third shape.	______	
	Draw a diamond in the rectangle.	______	
	Does not touch the side.	______	______
3.	Place a square above the star.	______	
	Cross out the truck.	______	
	Line under the car.	______	
	Two beneath the first object.	______	
	Circle around the fish.	______	______
4.	Draw a G inside the triangle.	______	
	Circle under the rectangle.	______	
	Line through the diamond.	______	______
5.	Cross out the circle.	______	
	Square around the small circle.	______	
	N in the blue square.	______	
	Underline the yellow square.	______	______
6.	Circle the hat.	______	
	Cross out the number ten.	______	______
7.	Draw a Z under the first square.	______	
	Underline the rectangle.	______	
	Cross out the last diamond.	______	
	Put a star on the second square.	______	______
8.	Put an S on the cow.	______	
	Circle the girl.	______	
9.	Connect the triangle and circle.	______	
	Cross out the star.	______	______
10.	Cross out the orange square.	______	
	Underline the sixth square.	______	
	Connect the yellow and green squares.	______	
	Circle the word free.	______	
	H between the words free and top.	______	______

11. Place the number ten in the triangle. ______
Cross out the house. ______
Draw a line under the cow. ______ ______

12. Connect the triangle and star. ______
Circle the diamond. ______
Four between the triangle and square. ______ ______

13. Underline the truck. ______
Cross out the apple. ______
Line connecting the tree and the glass. ______
LIne must go below the objects. ______
A star on top of the tree. ______ ______

14. Circle the pie. ______
Connect the square and the rectangle. ______
Draw a three under the house. ______ ______

15. Cross out the small star.
Square around the knife. ______
Square inside the rectangle. ______
Connect the biggest star to the car. ______ ______

16. Draw a diamond above the green circle. ______
Rectangle over the third circle. ______
Connect jail to fist. ______
With a line above the words. ______
Zero under the fourth word. ______ ______

17. Line between the orange and black shapes. ______
LIne goes above the shapes. ______
Star between the square and circle. ______
Circle the white rectangle. ______
Underline the diamond. ______ ______

18. U in the circle. ______
Underline the oval. ______
Triangle after the square. ______ ______

19. Circle the tree. ______
Line between the knife and pie. ______
A star in the square. ______ ______

20. Cross out the rectangle. ______
Letter B left of the triangle. ______ ______

21. Line between the small star and the rectangle. ______
Connect the stars. ______
Ccircle around the square. ______
Underline the fourth shape. ______ ______

22. Put a J between the fish and pie. ______
Circle the eight. ______ ______

23. Connect the first large and the second small star.		______
Underline the square.	______	
Circle the last two small stars.	______	
Three in the square.	______	______
24. Draw a line between the truck and boy.	______	
Circle the E.	______	______
Total/X/SD	______ ______	______

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Three.

			Total
1.	Connect the second and third circles.	______	
	A square around the word cat.	______	
	H on open.	______	
	Line under the number two.	______	______
2.	Four between the pink and yellow shapes.	______	
	Line through the triangle.	______	
	That touches the circle.	______	
	R to the right of the triangle.	______	
	Underline the square.	______	______
3.	Place the number three beside the girl.	______	
	Cross out the fork.	______	______
4.	Circle the letters inside the second rectangle.	______	
	Connect the first and second circle.	______	
	X above number fourteen.	______	
	B below thirty six.	______	______
5.	Connect the star with the small rectangle.	______	
	Square above the red shape.	______	
	Circle the oval.	______	
	Underline the blue shape.	______	______
6.	Write a D beside the ball.	______	
	Line around the spoon.	______	
	Cross out the man.	______	______
7.	Cross out the oval.	______	
	Connect the diamond to the triangle.	______	______
8.	Write a C in the large shape.	______	
	F below the pie.	______	
	Underline the diamond.	______	
	S square on top of the cow.	______	______
9.	Connect the large and small triangles.	______	
	With a line that goes above the shapes.	______	
	X between the purple and green shapes.	______	
	Underline the diamond.	______	
	Underline the oval.	______	______
10.	Make a square beside the rectangle.	______	
	Cross out the letter R.	______	______
11.	Circle the fork.	______	
	Square in front of the fish.	______	______

12. Two to the right of the car. ______
Connect the triangle to the pie. ______
Circle the star. ______
Underline the seven. ______ ______

13. Draw a line under the tree. ______
Cross out the house. ______
N in the diamond. ______ ______

14. Place an X below the circle. ______
Line from the orange diamond to the square. ______
That touches the blue diamond. ______
Eight in the rectangle. ______
Eight in the triangle. ______ ______

15. Connect the pink and black squares. ______
Circle the truck. ______
Underline the spoon. ______
Q to the left of the diamond. ______ ______

16. Make a square around the small circle. ______
That touches the sides. ______
Connect the white square and the diamond. ______
V in the oval. ______
Five in the circle. ______ ______

17. Draw a line between the two words. ______
Cross out the hat. ______
Square under the deer. ______ ______

18. Connect the fish and car. ______
Line under the rectangle. ______ ______

19. Cross out the knife. ______
Connect the tree to the apple. ______ ______

20. Put an O in the small square. ______
Cross out the large square. ______
Cross out the other large square ______ ______

21. Draw a square around the star. ______
X in the smallest shape. ______
That touches the side. ______
Connect the purple and green shapes. ______
E in the rectangle. ______ ______

22. Line between the tops of the blue and yellow shapes. ______
Diamond in the square. ______
Underline the first triangle. ______
and the circle. ______
Connect the circle and second triangle. ______ ______

23. Make a circle inside the diamond. _____
Connect the large and small rectangles. _____
Cross out the star. _____ _____

24. Make a K over the car. _____
P left of the apple. _____
Underline the tree. _____ _____

Total/x/SD _____ _____ _____

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Four.

No.	Instruction	Score	Total
1.	Cross out the biggest triangle.	______	
	R on top of the smallest.	______	
	X beside the white shape.	______	
	X inside the yellow.	______	______
2.	Line from the purple shape to the star.	______	
	That goes around the oval.	______	
	Nine left of the triangle.	______	
	Underline the square.	______	
	Underline the green shape.	______	______
3.	Cross out the apple.	______	
	Square next to the knife.	______	______
4.	Make a triangle next to the star.	______	
	Circle the car.	______	
	Connect it to the bottom of the pie.	______	
	Underline the number three.	______	______
5.	Connect the first and fourth square.	______	
	With a line that goes through number ninety.	______	
	K in the last square.	______	
	Cross out the NP.	______	
	Underline the twenty three.	______	______
6.	Write an O above the house.	______	
	Cross out the cow.	______	______
7.	Draw a square around the triangle.	______	
	Circle the P.	______	
	Connect the number eight and the diamond.	______	______
8.	Line from the purple shape to the D.	______	
	That goes under the C.	______	
	Underline the man.	______	
	X under the F.	______	
	X before the truck.	______	______
9.	Make a circle around the fourteen.	______	
	G on the diamond.	______	______
10.	Circle in the triangle.	______	
	That touches the sides.	______	
	Line from the pink to the orange shape.	______	
	That goes through the black.	______	
	Cross out the oval.	______	______

11. Connect the top of the house to the car.	______	
The cow to the deer.	______	
Circle the man.	______	______
12. Write an N next to the glass.	______	
Line beneath the tree.	______	
X over the hat.	______	
Connect the truck and apple.	______	______
13. Line from the star to the ball.	______	
That circles the pie.	______	
Cross out the tree.	______	______
14. Make a square around the fish.	______	
Five in the rectangle.	______	
X on the fork.	______	______
15. Underline the small circle.	______	
Connect the blue square with the second circle.	______	
Using a line over the shapes.	______	
Two in the small square.	______	
D on the green shape.	______	______
16. Cross out the oval.	______	
J in the rectangle.	______	______
17. Circle the big star.	______	
Connect it to the top of the knife.	______	
Underline the diamond.	______	______
18. Line from the small white shape to the star.	______	
That goes through the third shape.	______	
Connect the large and small rectangles.	______	
Underline the star.	______	
Cross out the oval.	______	______
19. L below the diamond.	______	
Three in the rectangle.	______	______
20. Square around pen.	______	
Circle around top.	______	
X left of the yellow circle.	______	
Q in the third circle.	______	______
21. Connect the spoon to the house.	______	
Underline the car.	______	
Two beneath the triangle.	______	______
22. Make a square around the black shape.	______	
Circle around the pink shape.	______	
Cross out the blue shape.	______	
Z in the oval.	______	______
23. Write a two left of the square.	______	
Underline the triangle.	______	______

24. X on the glass. ____
P on the blue shape. ____
Line under the spoon. ____
One right of the diamond. ____ ____

Total/x/SD ____ ____ ____

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Five.

1. Connect the small square and the glass. ______
   Underline the car. ______
   A five on the right of the fork. ______
   Cross out the large shape. ______ ______

2. Line between yellow shape & first small triangle. ______
   Circle the circle. ______
   X under the rectangle. ______
   Underline the purple square. ______
   Cross out the circle. ______ ______

3. Draw a square around the two. ______
   Connect the ten and six. ______ ______

4. Place an I in the green diamond. ______
   Connect the rectangle and the yellow shape. ______
   An oval between second diamond & the square. ______
   Circle the rectangle. ______ ______

5. Circle the fish. ______
   Q between the pie and the house. ______
   Cross out the cow's tail. ______ ______

6. Draw a star to the left of the girl. ______
   Connect the deer and car. ______
   Circle around the boy. ______
   Underline the cow. ______ ______

7. Make a V below the oval. ______
   Connect the star and small rectangle. ______
   Cross out the large rectangle. ______ ______

8. Circle the spoon. ______
   A line between the car and the spoon. ______ ______

9. Underline the first diamond. ______
   Connect the blue and pink shapes. ______
   B in the white diamond. ______
   Cross out the circle. ______ ______

10. Line between the orange and blue shapes. ______
    The line goes through the yellow circle. ______
    Underline the green shape. ______
    Cross out the white and middle circles. ______ ______

11. Connect the first and last circle. ______
    L under cat. ______
    A seven above cat. ______
    A star over door. ______
    Underline jail. ______ ______

12. Cross out the star. ______
A zero in the rectangle. ______

13. Connect the large circle and the small square. ______
Underline the oval. ______
Cross out the diamond. ______ ______

14. Draw a circle around the six. ______
Underline the glass. ______
Connect the fish's tail and the hat. ______ ______

15. Draw a line between the purple shapes. ______
X under the first shape. ______
X under last shape. ______
A circle between the square and diamond. ______ ______

16. Circle the apple. ______
A C to the right of the deer. ______ ______

17. Underline the smallest shape. ______
A star under the orange shape. ______
Cross out the oval. ______
An F in the rectangle. ______ ______

18. A ten above the small circle. ______
Underline the last shape. ______
Connect the yellow and orange shapes. ______
A diamond around the first shape. ______
Cross out the square. ______ ______

19. Connect the ball and the spoon. ______
A square under the car. ______ ______

20. Cross out the square. ______
An R above number twenty three. ______

21. Connect the first and last words. ______
A W above the girl. ______
Cross out the car. ______
An oval to the left of the house. ______
Circle the second object. ______ ______

22. A J behind the fish. ______
Underline the rectangle. ______
A line between the star and triangle. ______ ______

23. Connect the small star and diamond. ______
Underline the orange shape. ______
Cross out the yellow shape. ______
A C above and below the rectangle. ______
Circle the star. ______ ______

24. A K on the truck. _____
Underline the fork. _____
The number three below the hat. _____ _____

Total/X/SD _____ _____ _____

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Six.

			Total
1.	A D in the triangle.	______	
	A number four in the diamond.	______	
	Underline the rectangle.	______	______
2.	Draw a line from door to bus.	______	
	Underline cat.	______	
	A U in the shape.	______	______
3.	Underline the first word.	______	
	Connect the ball to the second word.	______	
	A line beside the hat.	______	
	A G over the deer.	______	______
4.	Make a square around the house.	______	
	A circle around the boy.	______	
	A nine below the car.	______	
	Connect the third and fourth objects.	______	______
5.	Put an X under the fish's tail.	______	
	Cross out the pie.	______	______
6.	Connect the two black shapes.	______	
	With a line that circles the square.	______	
	An M in the rectangle.	______	
	Underline the triangle.	______	
7.	Draw a line through the Y.	______	
	Connect the N to the S.	______	
	Cross out the U.	______	______
8.	Circle the oval.	______	
	A two in the triangle.	______	______
9.	A square around the third big star.	______	
	Connect it to the second small star.	______	
	Circle the number.	______	
	Underline the square.	______	
	Cross out the first star.	______	______
10.	Write a four in the square.	______	
	X over the girl.	______	
	A two below the apple.	______	______
11.	A square around the yellow shape.	______	
	That does not touch it.	______	
	Z left of the second shape.	______	
	A five above the green shape.	______	
	Underline the diamond.	______	______

12. Connect the orange shape to the white rectangle. ______
An R between the hat and the fork. ______
A square behind the fish's tail. ______
Connect all the shapes ______ ______

13. Draw a line beside letter A. ______
A line through E. ______ ______

14. An eight in the second triangle. ______
A four in the orange shape. ______
Connect the yellow triangle with the square. ______
With a line over the shapes. ______
Underline the circle. ______ ______

15. Make a triangle around the deer. ______
A line through the fork. ______
X right of the apple. ______ ______

16. Circle number thirty six. ______
A square around number one. ______ ______

17. A J in the oval. ______
A circle over the rectangle. ______
A line through the triangle. ______ ______

18. Connect open to can. ______
With a line that circles the oval. ______
Cross out the two. ______
A square around the second circle. ______ ______
A diamond to the left of cat.

19. A B in the green shape. ______
A five in the small white shape. ______
Circle the triangle. ______
Cross out the orange rectangle. ______ ______

20. A line from the blue square to yellow diamond. ______
It does not touch another shape. ______
Cross out the biggest rectangle. ______
Circle the oval and underline the star. ______ ______

21. Connect the spoon to the apple. ______
Circle the glass. ______ ______

22. Cross out the knife. ______
An M under the tree. ______ ______

23. An X in the orange shape. ______
An E in the diamond. ______
Circle the square. ______
Connect the two ovals. ______
A line that goes through the rectangle. ______ ______

24. Circle the purple triangle. _____
A square between the green and blue shapes. _____
An X over the star. _____
A inside the rectangle. _____ _____

Total/X/SD _____ _____ _____

Appendix F. (continued)

Multiple Instructions Test. Score Form Number Eight

		Total
1. A line from the two to the six.	_____	
Connect the second rectangle to the knife.	_____	
Underline the two.	_____	
Circle the hat.	_____	_____
2. A square under the purple shape.	_____	
A G beside the oval.	_____	
A line between the orange and yellow shapes.	_____	
A line around the star.	_____	_____
3. Circle the AB.	_____	
Connect it to the NP.	_____	
A line between thirty six and fifty seven.	_____	_____
4. A five in the circle.	_____	
A three below the diamond.	_____	
A line between the pink and the green shapes.	_____	
The line goes through the blue.	_____	
Cross out the rectangle.	_____	_____
5. An F left of the apple.	_____	
A T on the glass.	_____	
A circle around the woman.	_____	
Underline the tree.	_____	_____
6. A seven between the stars.	_____	
Cross out the car.	_____	_____
7. Cross out the smallest shape.	_____	
Connect the pink and the yellow shapes.	_____	
Without touching any others.	_____	
An L in front of the cow.	_____	
A four over the house.	_____	_____
8. Connect third and fourth object.	_____	
An eight beneath back of deer.	_____	
A line through ball.	_____	_____
9. An X on the pie.	_____	
Circle the glass.	_____	_____
10. A six in the rectangle.	_____	
A B in the diamond.	_____	
A ten in the triangle.	_____	_____
11. A line over the number nine.	_____	
A circle around the number fourteen.	_____	
Connect the third and fifth numbers.	_____	
Cross out number fifty seven.	_____	_____

12. Connect black square to second small triangle. ______
A line going around the purple triangle. ______
Underline the small circle. ______
A Q in the rectangle. ______ ______

13. A rectangle around the pie. ______
A line from the star to the seven. ______
The line goes over the ball. ______
A J behind the car. ______
A P under the triangle. ______ ______

14. Cross out the fork. ______
A four before the truck. ______
An X next to the girl. ______ ______

15. Circle the word free. ______
Connect it to the third square. ______
A C in the second. ______
A nine in the purple square. ______
Connect the words bus and top. ______ ______

16. A line over the circle. ______
Cross out the triangle. ______ ______

17. An X on the ball. ______
An eight in front of the deer. ______
Underline the apple. ______ ______

18. A line from the first shape to the yellow. ______
That goes under the shapes. ______
A circle under the small rectangle. ______
A star under the purple. ______
A D in the oval. ______ ______

19. Circle the F. ______
Connect the man to the truck. ______ ______

20. Make a circle around the diamond. ______
Cross out the oval. ______ ______

21. A two in the square. ______
An L in the rectangle. ______
A line over the star. ______ ______

22. Connect the triangle to the square. ______
The oval to the diamond. ______ ______

23. A line between the blue and the orange shapes. ______
A line through the two. ______
Underline can. ______
Cross out the second number. ______ ______

24. Make a square around the house. _____
An X beside the ball. _____
A circle beneath the apple. _____ _____

Total/X/SD _____ _____ _____

REFERENCES

References

AMERICAN NATIONAL STANDARDS INSTITUTE, American National Standards for Specifications for Audiometers. ANSI S3.6-1969. New York: American National Standards Institute (1969).

Anderson, H., Barr, B., and Wedenberg, E. (1970). Early diagnosis of the eighth-nerve tumors by acoustic reflex tests. ACTA Otolaryngologica Supplement 263, 232-237 (1970).

Berger, K. (1978). Speech audiometry. In D. Rose (Ed.), Audiological Assessment. Englewood Cliffs, N. J.: Prentice Hall.

Bernstein, J., and Pisoni, D. B. (1981). Unlimited text-to-speech device; Description and evaluation of a microprocessor-based system. In 1980 IEEE International Conference Record on Acoustics, Speech and Signal Processing (pp.576-579). New York. IEEE Press.

Bilger, R. C. (1983). Manual for the clinical use of the revised SPIN test.

Bilger, R. C. Nuetzel, J. M., Ranbinowitz, W. M., and Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. Journal of Speech and Hearing Research. 27:32-48.

Beukelman, D. R., Yorkston, K. M., Popete, M. and Naranjo, C. (1984). Frequency of word occurrence in communication samples produced by adult communication aid users. Journal of Speech an Hearing Disorders. 49,360-367.

Chial, M. R. (1973). The masked intelligibility threshold test for speech communication system evaluation. Unpublished Doctoral Dissertation, University of Wisconsin, Madison.

Chial, M. R. (1976, April). Evaluation of a synthetic talker for speech intelligibility testing. Poster Presentation G9 at the Minty-First Meeting of the Acoustical Society of America, Washington, D. C..

Chial, M. R. (1984). Evaluating microcomputer hardware. In Schwartz, A. H. (Ed) Handbook of Microcomputer Applications in Communication Disorders. San Diago, California: College-Hill Press, 275-310.

Chial, M. R. (1986, May).Tutorial on Speech Synthesis. Presentation at Northwestern University, Evanston, Ill.

Chial, M. R. and Daniel, S. W.(November, 1977).Hearing aid quality judgements by normal and dysacusic listeners. Paper presented at the Annual Convention of at the American Speech and Hearing Association, Chicago.

Chial, M. R., and Hayes, C. S.(1974).Hearing aid evaluation methods: Some underlaying assumptions. Journal of Speech and Hearing Disorders, 19.270-279.

Connors, K. M. (1974). The effect of final pause duration and vowel duration upon the comprehension and delayed recall of oral prose. Unpublished Masters Thesis, The University of Texas, Austin, Texas.

Cox, R.M. and McDaniel, D.M. (1984). Intelligibility of continuous discourse: application to hearing aid selection. Journal of the Acoustical Society of America, 76:758-766.

Dirks, D.D., Kamm, C. A., Dubno, J. R. and Velde, T. M. (1981). Speech recognition performance at loudness discomfort level. Scandinavian Audiology, 10:239-246.

Dirks, D. D., Morgan, D. E., and Dubno, J. R. (1982). A procedure for quantifying the effects of noise on speech recognition. Journal of Speech and Hearing Disorders, 47:114-123.

Dubno, J. R., Dirks, D. D. and Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. Journal of the Acoustical Society of America, 76:87-96.

Dvorine, I. (1953). Dvorine Pseudo-Isochromatic Plates, The Psychological Corporation.

Fairbanks, G., Guttman, N., and Miron, M. S. (1957). The effects of time compression upon the comprehension of connected speech. Journal of Speech and Hearing Disorders, 22, 10-19.

Flanagan, J. L. (1981). Synthesis and recognition of speech: How computers talk, Bell Laboratories Record. April:123-130.

Greene, B. G., Logan, J. S., and Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: intelligibility of eight test-to-speech systems. Behavior Research Methods, Instruments, and Computers, 18:100-107.

Greene, B.G. and Pisoni, D.B. (1988). Perception of synthetic speech by adults and children:research on processing voice out put form test-to-speech systems. In The Vocally Impaired: Clinical Practice and Research. Lynne E. Berstein. (Ed) Philidelphia: Grune & Stratton

Greene, B. G., Manous, L. M., and Pisoni, D. B. (1984). Perceptual evaluation of DECtalk; Final report of version 1.8 Research of Speech Perception Progress Report No. 10. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology

Greenspan, S., Nusbaum, H. C., and Pisoni, D. B. (1984). Perception of synthetic speech: Some effects of training and attentional limitations. Research on Speech Perception Progress Report No. 11. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Greenspan, S., Nusbaum, H. C., and Pisoni, D. B. (1986). Perceptual learning of synthetic speech produced by rule. Research on Speech Perception Progress Report No. 12. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Gough, P. B. (1956) Grammatical transformations and speed of understanding.Journal of Verbal Learning and Verbal Behavior. 4,107-111.

Gough, P. B. (1966).The verification of sentences: The effects of delay of evidence and sentence length. Journal of Verbal Learning and Verbal Behavior. 5,492-496.

Gross, E. (1966). Thunder: The Writers Assistant. Richmond Hill; Ontario.

Hammill, D. D. (1985). Detroit Tests of Learning Aptitude. Pro-ed. Austin: Texas.

Heidinger, V. A. (1984). Analyzing Syntax and Semantics. Washington, D.C; Gallaudet College Press.

Hersch, H. M., and Tartaglia, L. (1983). Understanding synthetic speech. User Research Group, Corp. Res. and Architecture, Maynard, Mass: Digital Group.

Hoover, J., Reichle, J. Van Tasell, D. and Cole, D. (1987). The intelligibility of synthesized speech: Echo II versus Votrax. Journal of Speech and Hearing Research. 30:425-431.

Huggins, A. W. F., and Nickerson, R. S. (1985). Speech quality evaluation using "phoneme-specific" sentences. Journal of the Acoustical Society of America, 77:1896-1906.

Hutchenson, R. W., Dirks, D. D., and Morgan, D. E. (1979). Evaluation of the speech perception in noise (SPIN) test. Otolaryngology Head Neck Surgery, 87:239-245.

IEEE Standards Committee, Rothaure, H. E., Chairman. IEEE Recommended Practice for Speech Quality Measurements. IEEE Transactions on Audio and Electronics, 17,225-246.

Jerger, J. (1970). Clinical experience with impedance audiometry. Archives of Otolayrngology, 92, 311-324.

Jenkins, J. J. and Franklin, L. D. (November,1981). Recall of passages of synthetic speech. Paper presented at the Psychonomics Society Meeting.

Kalikow, D.N., Stevens, K.N., and Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. Journal of the Acoustical Society of America, 61:1337-1351.

Klatt, D.H. (1987). Review of test-to speech conversion for English. Journal of the Acoustical Society of America, 82:737-792.

Kraat, A. and Levinson, E. (October, 1984). Intelligibility of two speech synthesizers used in augmentatve communication devices for severely speech impaired. Paper presented at the Third International Conference on Augmentative and Alternative Communication. Massachusetts Institute of Technology.

Larkey, J.S. and Danly, M. (1983). Fundamental frequency and sentence comprehension. MIT Speech Group Working Papers, Vol II.

Lawson, G. D. and Chial, M.R. (1982). Quality magnitude estimation of degraded speech by listeners with normal- and impaired-hearing listeners. Journal of the Acoustical Society of America, 72:1781-1787.

Leavitt, H. J. and Mueller, R. H. (1951). Some effects of feedback on communication. Human Relations. 4,401-410.

Levinson, S. E., and Liberman, M. Y. (1981). Speech Recognition by computer. Scientific American, April:64-76.

Lehiste, I. and Peterson, G. E. (1961). Transitions, glides, and dipthongs. Journal of the Acoustical Society of America. 33:268-277.

Logan, J. S.,and Pisoni, D. B. (1986). Intelligibility of phoneme specific sentences using three text-to-speech systems and a natural speech control. Research Perception Progress Report No. 12. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Logan, J. S., Pisoni, D. B., and Greene, D.B. (1985). Measuring the segmental intelligibility of synthetic speech; MRT results for eight text-to-speech systems. Research Perception Progress Report No. 11. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Manous, L. M., Pisoni, D. B., Dedina, M. J., and Nusbaum, H. C. (1985).Comprehension of natural and synthetic speech using a sentence verification task. Research on Speech Perception Progress Report No. 11. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

McCandless, G. A., and Thomas, G. K. (1974). Impedance audiometry as a screening procedure for middle ear disease.Transripts of the American Academy of Opthalmology and Otolaryngology. 78, 98-102.

McHugh, A. (1976). Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Navel Research Lab (9-9-76)

Miller, J. D. (1982).A phonetically relevent auditory-perceptual space.Journal of the Acoustical Society of America. Suppl. 1., 72:S64.

Miller, J., D.(1984). Implications of the auditory-perceptual theory of phonetic perception for speech recognition by the hearing impaired. ASHA Reports. 14:45-48.

Mirenda, P.and Beukelman, D. R.(1987). A comparison of speech synthesis intelligibility with listeners from three age groups. Augmentative and Alternative Communication. 120 -128.

Morgan, D. E., Kamm, C. A., and Velde, T. M. (1981). Form equivalence of the speech perception in noise (SPIN). Journal of the Acoustical Society of America, 69:1791-1798.

Munson, W. A., and Karlin, J. E. (1962). Isopreference method for evaluating speech transmission circuits. Journal of the Acoustical Society of America. 34, 762-774.

Nixon, C.W., Anderson, T.R. and Moore, J.J. (1985). The perception of synthetic speech in noise. Armstrong Aerospace Medical Research Laboratories Report. Wright-Patterson AFB, OH 45433.

Northern, J.L. and Grimes, A. M. (1978). Introduction to acoustic impedance. In J. Katz (Ed.), Handbook of Clinical Audiology (pp. 344-355).Baltimore: The Williams & Wilkins Company.

Nusbaum, H.C., Greenspan, S. L., and Pisoni, D. B.. (1986). Perceptual attention in monitoring natural and synthetic speech. Research Perception Progress Report No. 12. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Nusbaum, H.C.,Schwab, E. C. and Pisoni, D. B.(1984).Subjective evaluation of synthetic speech: Measuring preference, naturalness, and acceptability. Research on Speech Perception Progress Report No. 10, Bloomington, IN: Speech Research Laboratory, Indiana University.

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

Nye, P. W., and Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. Haskins Laboratories Status Report on Speech Research, 38, 169-190.

Nye, P. W., Ingemann, F., and Donald, L. (1975). Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. Haskin Laboratories: Status Report on Speech Perception SR-41.

Oyer, H. J. and Frankman, J. P.(1975). The Aural Rehabilitation Process, A Conceptual Framework Analysis. New York: Holt, Rinehart and Winston.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of vowels. Journal fo the Acoustical Society of America, 24:173-184.

Pisoni, D. B. and Dedina, M. J. (1986). Comprehension of digitally encoded natural speech using a sentence verification task (SVT): a first report. Research on Speech perception Progress Report No. 12. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Pisoni, D. B., and Hunnicutt, S. (1980). Perceptual evaluation of MITtalk; The MIT unrestricted text-to-speech system. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (pp.572-575). New York. IEEE Press.

Pisoni, D. B., & Koen, E. (1981). Some comparisons of intelligibility of synthetic and natural speech at different speech-to-noise ratios. Research on Speech perception Progress Report No. 7. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Pisoni, D. B., Manous, L. M. and Dedina, M. J. (1986). Comprehension of natural and synthetic speech: II. Effects of predictability on the verification of sentences controlled for intelligibility. Research on Speech perception Progress Report No. 12. Bloomington Indiana University, Speech Research Laboratory, Department of Psychology.

Pisoni, D. B., Nusbaum, H. C., & Greene, B. G. (1985). Perception of synthetic speech generated by rule. Proceedings of the IEEE, 73:1665-1676.

Rosenberg, S. and Hall, R. L. (1958). The effects of different social feedback conditions upon performance in dyadic teams. Journal of Abnormal and Social Psychology, 57, 271-277.

Rothauser, E. H., Urbanek, G. E., and Pachl, W. P. (1968). Isopreference methods for speech evaluation. Journal of the Acoustical Society of America, 44, 408-418.

Schiavetti, N., Sitler, R., Metz, D. E., & Houde, R. A. (1984). Prediction of contextual speech intelligibility from isolated word intelligibility measures. Journal of Speech and Hearing Research, 27:623-626.

Schwab, e. C.,Nusbaum, H.C. and Pisoni, D.B.(1985).Some effects of training on the perception of synthetic speech. Human Factors, 27,

Shannon, C. E. and Weaver, W. (1949). The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Shepard, R.N. (1972). Psychological representation of speech sounds. In E.E. David and Denes (Eds.), Human Communication: A Unified View (pp. 67-113). New York: McGraw-Hill.

Simpson, A.C. and Williams, D.H. (1980). Response time effects of alerting tone and semantic context for synthesized voice cockpit warnings. Human Factors, 22,319-330.

Slowiaczek, L.M. and Pisoni, D. B. (1982). Effects of practice on speeded classification of natural and synthetic speech. Journal of the Acoustical Society of America, 71, S95-96.

Smith, A. G. (1966). Feedback and control. In A. G. Smith (Ed.) Communication and Culture. New York: Holt, Rinehart and Winston, 322-323.

Studebaker, G.A. (1985). A "rationalized" arcsine transform. Journal of Speech and Hearing Research. 28,455-462.

Thordike, E.L.,and Lorge,I. (1944).The Teachers Book of 30,000 Words. New York: Teacher's College, Columbus University,1944.

Weldele, F. S. and Millin, J. P. (1975). The usefulness of quantitative listener judgements in evaluating hearing aids. Ohio Journal of Speech and Hearing, 11:39-47.

Yorkston, K. M., and Beukelman, D. R., (1981). Assessment of Intelligibility of Dysarthric Speech. Seattle: C.C.Publications Inc.

Yorkston, K. M., Beukelman, D. R., and Traynor, C. (1984). Computerized Assessment of Intelligibility of Dysarthric Speech, Austin: Pro-Ed.