DOCUMENT RESUME

ED 041 477 EM 008 200

AUTHOR	Utz, Walter J., Jr.
TITLE	The Use of Computer Generated Tests to Select a Speaker for a Random Access Digital Audio System.
INSTITUTION	Radio Corp. of America, Palo Alto, Calif. Instructional Systems.
PUB DATE	Apr 70
NOTE	12p.; Paper presented at Annual Meeting of the Department of Audio-Visual Instruction, National Education Association (Detroit, Michigan, April 27 - May 1, 1970)
AVAILABLE FROM	RCA Instructional Systems, 530 University Avenue, Palo Alto, California 94301 (copies of tapes)
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.70
DESCRIPTORS	Articulation (Speech), *Artificial Speech, Attitude Tests, Audio Equipment, *Computer Assisted Instruction, *Listening Comprehension

ABSTRACT

Computerized speech could enhance the effectiveness of computer-assisted instruction as an educational tool. Digital audio under computer control allows a very wide range of replies, but it poses special problems in the areas of listener attitudes and speaker intelligibility. This paper discusses the design and implementation of special tests to discover a speaker who would be most pleasing and intelligible to students using a random access digital audio in a computer-assisted instruction system. Auditions were for both amateur and professional speakers, male and female. Junior college students rated the voices for likeability and intelligibility. Those who scored highest in the two tests all had some professional voice training and spoke in a mid-range pitch. As was expected , there was a correlation between intelligibility and attitude. Appendices contain raw scores and illustrative figures. (JY)

Walter J. Utz, Jr.

THE USE OF COMPUTER GENERATED TESTS TO SELECT A SPEAKER FOR A RANDOM ACCESS DIGITAL AUDIO SYSTEM

The spoken word is an integral part of a child's education, and computerized speech could enhance the effectiveness of computer-assisted instruction as an educational tool. Conventional analog tape recording methods do not readily permit random access of numerous replies to cover a wide range of learning situations. Digital audio under computer control allows a very wide range of replies, but it poses special problems in the areas of listener attitude and speaker intelligibility. This paper will discuss the design and implementation of special tests to discover a speaker who would be most pleasing and most intelligible to students using random access digital audio in our computer-assisted instruction system.

Let us begin with an examination of the basic difference between analog and digital audio. Figure 1 shows one of the many methods we have to store sounds; in this case, by musical notation. The listener, a trained musician, converts the musical tones that he hears to musical notes which he records on paper. In this written form the music can be stored indefinitely, but it can be reproduced as music at any time by another trained musician.

Another storage system, the most efficient way to store sounds for computer control, is to convert sound's analog signal into a digital format for computer processing as shown in Figure 2. The digital format permits an ease of access and control for the audio information, and it also permits storage on a standard computer disc unit.

For those of you who are not familiar with a computer disc unit, one is shown in Figure 3. Note the similarity to record discs. These discs are coated with a magnetic recording substance which may be reached by the movable heads shown to your left. The important thing to be known here is that there are 2000 recording tracks on such a unit, and any track can be reached in less than one-tenth of a second. Digital audio stored on these tracks may be accessed quickly to compose sentences for playback as shown in Figure 4.

Although intelligible speech has been synthesized by various methods, the artificial speech quality has been judged to be a possible source of interference with the learning process at this stage of synthesized speech development. Thus we have chosen to operate at the word level, with sentences constructed from whole words that have previously been stored on a computer disc unit. This would be approx- the same as recording several thousand words on small lengths of recording tape, and then composing a message by splicing the proper pieces of tape. The computer performs the task at the rate of approximately 40 words per second, and this permits the composition of messages for more than one user at a time.

The tape splicing or computer splicing of words to form sentences leads to the first problem in the area of learning. The message must be understandable, and yet it is being composed of words spoken out of context. The speaker who is chosen for such a digital audio system must be able to pronounce the words in such a way as to minimize the contextual conflicts in pronunciation while at the same time achieving a high rate of intelligibility. In this case intelligibility is the prime factor with attitude playing a major supporting role.

The ability to achieve a high rate of intelligibility while minimizing the contextual problem of pronunciation might not be restricted to professional announcers. Our auditions included both amateur and professional speakers with approximately an equal number of males and females. Each speaker read a list of monosyllables chosen at random from the Harvard monosyllable lists, and they also read sentences designed to cover the normal range of pronunciation problems.

The time and effort required to run intelligibility tests dictated of necessity our decision to run the attitude tests first, and then measure the intelligibility levels of the top seven speakers. The test design is a balanced incomplete factorial design as shown in Figure 5. In this test, every speaker is compared to every other speaker twice to permit each speaker to have the first position in a binary comparison. The test is divided into many subsections in which the listeners hear one speaker and then another. The listeners are then asked to indicate their preference for speaker A, speaker B, or neither speaker. There are 342 speaker comparisons, and each test group (there are six groups) is asked to rate one-sixth of the comparisons, or 57.

Each comparison consists of one speaker saying three words, and then another speaker saying the same three words. To eliminate listener fatigue, there are ten words in a list, and each comparison moves to the next three words on the list. Thus the speakers and words are constantly changing. To produce the type of test I have just described by conventional tape splicing or dubbing methods would be a considerable effort. The audio delivery program was modified to have the computer select the six words for each pair of speaker comparisons and the test tapes were produced under computer control in less than two hours. Note that the computer not only selected the word pairs, it also played the audio comparisons. Then a regular tape recorder was used to record the audio test generated by the computer. Here is a sample of the comparison tapes; all nineteen voices are included in the sample. (Play audio tape segment one).

The seven finalists with the highest scores in the attitude test were allowed to read the intelligibility tests, which are constructed from six standard intelligibility tests as specified by the Acoustical Society of America.[1] Each test contains 50 monosyllabic words, and each word is spoken in the statement "Would you write _____ now?" read as a simple declarative sentence. In this case the computer was not used; rather a delta modulation simulator was used to provide the equivalent audio output for the intelligibility tests. The computer could have been employed to generate the tests, but the linear nature of the material permitted a straightforward recording approach. Here are recorded samples of the seven speakers who participated in the intelligibility tests. (Play audio tape segment two).

The tests were administered to the six listener groups over a two day interval in the same room with the same playback configuration. The listeners wore stereo earphones which were connected in a monaural mode. Foothill Junior College students were paid for their participation, and they were selected on the basis of their willingness to participate. Any hearing defect automatically disqualified a potential test subject.

The tests went well. The students were generally eager to participate, and they definitely had opinions about the speakers, as the test results show. The test design had been pretested on a group of randomly selected RCA employees, and this helped to eliminate any potential confusion in the real tests. At least two persons were present to supervise each group of six students, and ensure that no horseplay or confusion arose.

The tests were graded by two independent groups to ensure accuracy. The attitude scores are shown in Figure 6. The adjusted score is obtained by adding two points for each win and one point for each tie. The top two scores have a considerable margin over the next six scores which are in the 220-230 range. Also, note that the top score is greater than three times the smallest score.

The intelligibility scores are shown in Figure 7. Although the same speaker scored highest in both test phases, there is a change in the second highest position. Speaker O, a commercial radio announcer, has an 88% intelligibility score, although he is more than 40 points lower in attitude than speaker F.

The four highest scoring speakers had some form of professional speech training, and one is a commercial radio announcer in San Francisco. In general, the female voices tend to be low in pitch while the male voices tend to be high among the high scorers. This would suggest

[1] American Standard Method for Measurement of Monosyllabic Word Intelligibility, Sponsored by the Acoustical Society of America. Approved May 25, 1960.

that a mid-range pitch might be best for our digital audio system. Note the consistency in the attitude and intelligibility scores. There may be an interaction at work here as a high intelligibility score may produce a high attitude rank. One important feature of liking a voice should be understanding the voice.

The highest scoring voice was used to produce a working dictionary of approximately 600 words to be used for a digital audio system as part of a computer-assisted instruction system. Here are some computer output. (Play audio tape segment three). Although it will probably never be possible to reproduce perfectly natural speech from words spoken out of context, the sample you have just heard is well over 90% intelligible when played over earphones in our installation.

Future studies should be performed to determine the type of voice best suited to a learning situation, or if many voices will serve in this application. The listener fatigue effect should be studied to see if digital audio becomes more or less pleasant with time. And in all of these studies it should be possible to use the computer to generate many tests in a fraction of the time necessary with analog recording techniques. The quality of digital audio is a function of the storage space required on the disc unit. If fewer words are stored, the quality of the digital audio system can be greatly enhanced while the advantages of computer processing are retained.

Further research in synthesized speech may permit us to generate thousands of words from some type of basic speech units. In the meantime we are striving to produce the best possible word oriented system to be used in industrial and computer-assisted instruction applications.

FIGURE 6

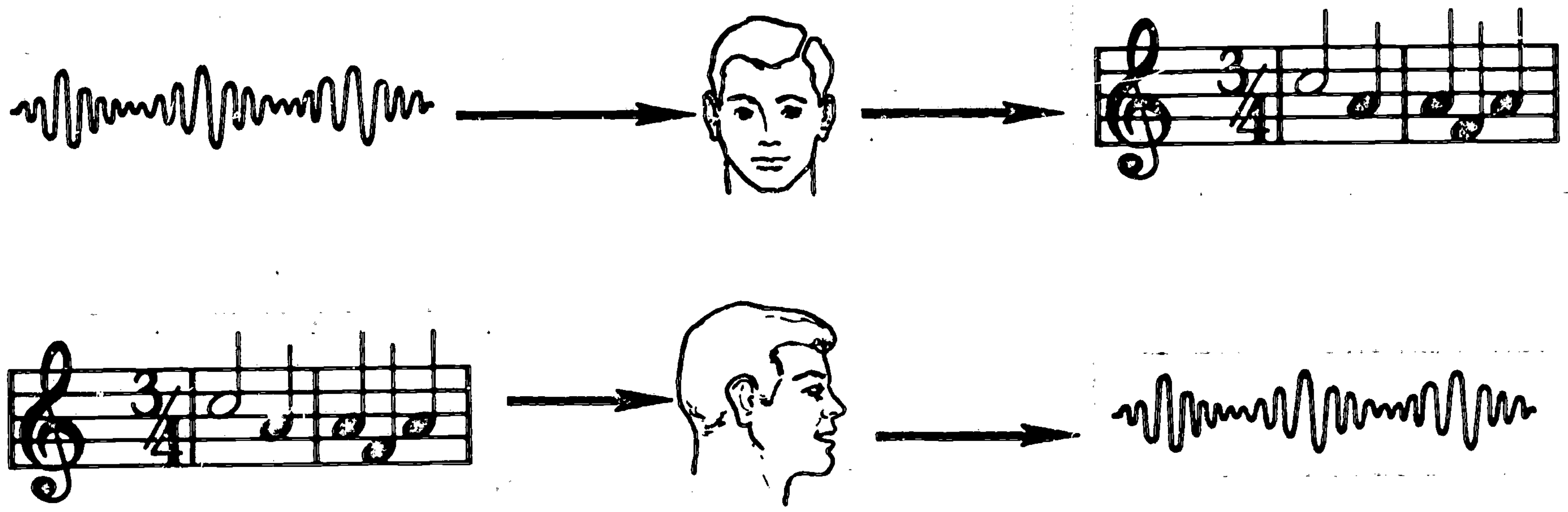
Phase I Attitude Scores

Speaker	Wins	Ties	Adjusted Scores
I	127	45	299
F	107	54	268
E	96	36	228
C	90	46	226
J	92	41	225
P	91	42	224
O	93	37	223
L	92	36	220
A	84	41	209
G	85	31	201
N	82	26	190
H	74	36	184
R	66	39	171
S	64	31	169
B	67	44	168
M	62	34	158
Q	57	35	149
D	63	13	139
K	38	14	90

FIGURE 7

Phase II Intelligibility Scores

	Intelligibility		
Speaker	Raw Score	Percent	Attitude Scores
I	734	90	299
O	730	88	223
J	688	84	225
F	682	83	268
C	652	80	226
E	576	70	228
P	503	61	224



F I G U R E   1

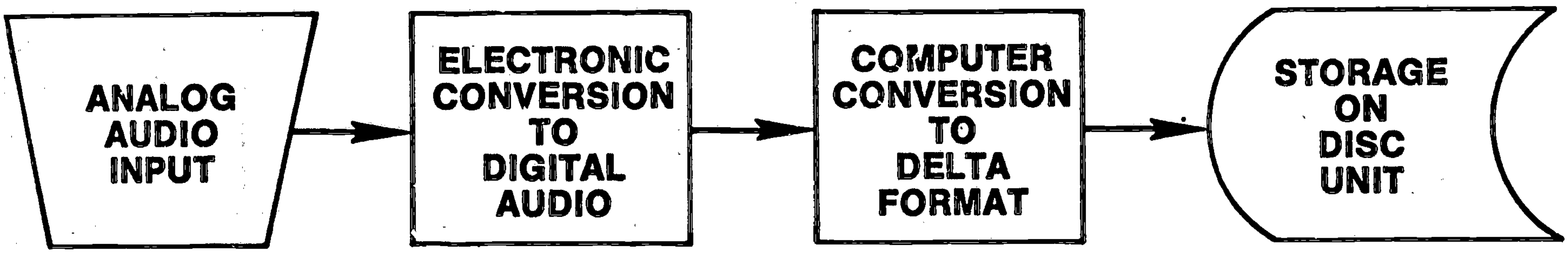
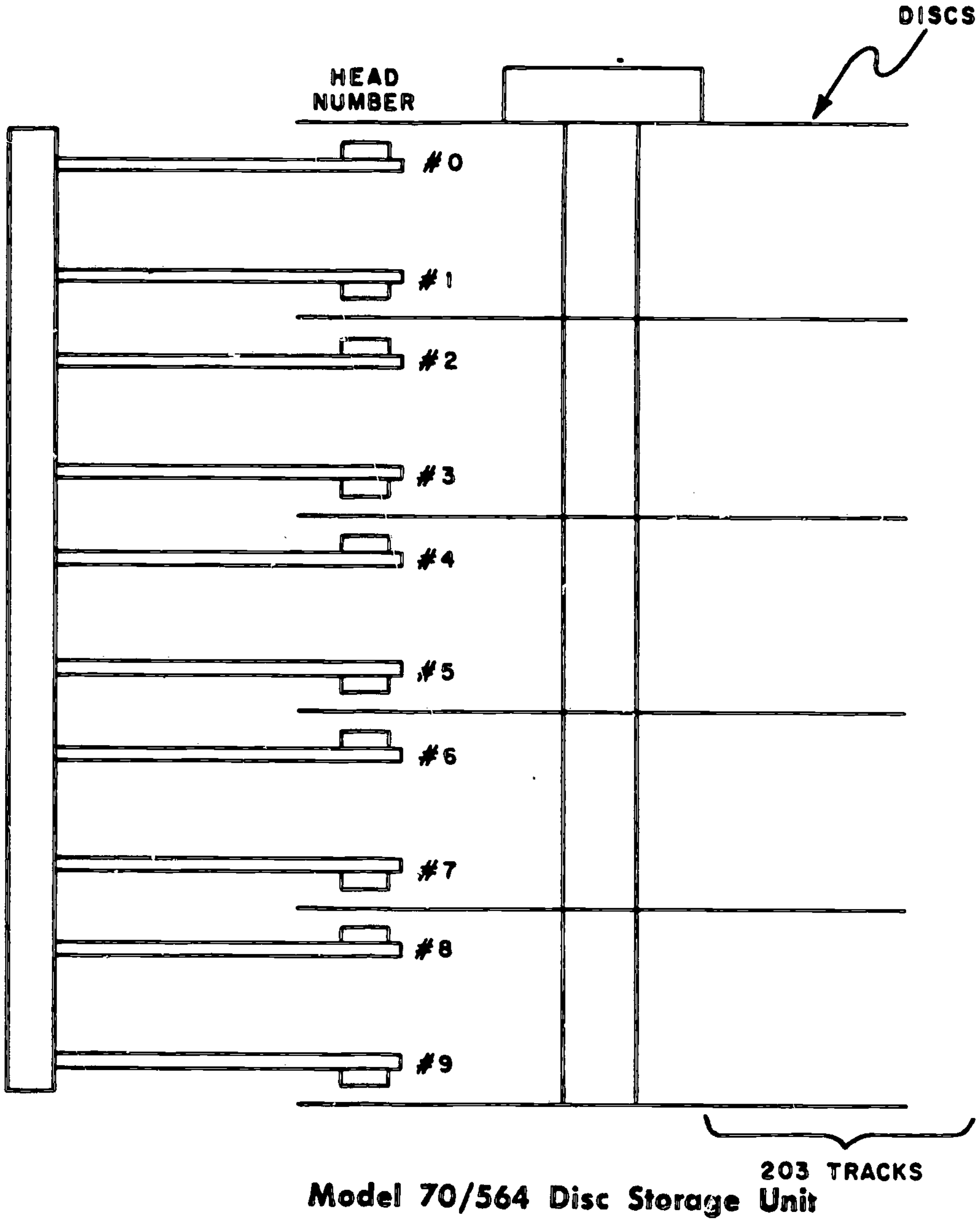

ANALOG AUDIO INPUT
ELECTRONIC CONVERSION TO DIGITAL AUDIO
COMPUTER CONVERSION TO DELTA FORMAT
STORAGE ON DISC UNIT

FIGURE 2

Cylinder
(Cont'd)

Model 70/551 Random Access Controller



Model 70/564 Disc Storage Unit

FIGURE 3

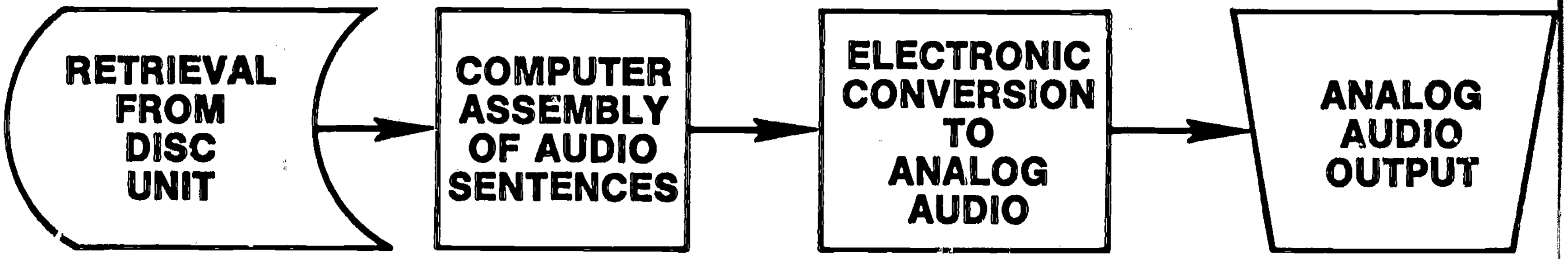

RETRIEVAL FROM DISC UNIT
COMPUTER ASSEMBLY OF AUDIO SENTENCES
ELECTRONIC CONVERSION TO ANALOG AUDIO
ANALOG AUDIO OUTPUT

FIGURE 4

GROUP

1	2	3	4	5	6
BA1	FD2	CS3	EA4	LG5	PJ6
DC4	IG5	LI6	FB7	NI8	RL9
FE7	LJ8	PM9	GC10	PK1	AN2
HG10	OM1	AQ2	HD3	RM4	DQ5
JI3	RP4	KB5	IE6	AO7	FS8
LK6	BS7	OF8	JF9	CQ10	OC1
NM9	KC10	SJ1	KG2	ES3	BI4
PO2	NF3	DN4	LH5	MB6	EL7
RQ5	QI6	HR7	MI8	OD9	GN10
AS8	AL9	QB10	NJ1	QF2	JQ3
IB1	DO2	CG3	OK4	SH5	LS6
KD4	GR5	FJ6	PL7	BJ8	AB9
MF7	PB8	KO9	QM10	DL1	CD2
OH10	SE1	NR2	RN3	FN4	EF5
QJ3	CH4	DA5	SO6	HP7	GH8
SL6	FK7	IF8	AP9	JR10	IJ1
BN9	IN10	MJ1	BQ2	AC3	KL4
DP2	LQ3	QN4	CR5	DF6	MN7
FR5	CA6	BR7	DS8	GI9	OP10
NA8	GE9	LC10	KA1	KM2	QR3
PC1	JH2	PG3	LB4	NP5	GA6
RE4	MK5	AK6	MC7	QS8	IC9
AG7	PN8	EO9	ND10	FA1	KE2
CI10	SQ1	IS2	OE3	HC4	NH5
EK3	IA4	RC5	PF6	JE7	SM8
GM6	LD7	BF8	QG9	SN10	NB1
IO9	OG10	GK1	RH2	BP3	PD4
KQ2	RJ3	JN4	SI5	DR6	QE7
MS5	BM6	MQ7	AJ8	LA9	CJ10
CB8	EP9	EB10	BK1	NC2	DK3
ED1	HS2	GD3	CL4	PE5	FM6
GF4	QC5	KH6	DM7	RG8	HO9
IH7	AF8	OL9	EN10	AI1	SA2
KJ10	DI1	SP2	FO3	CK4	BC5
ML3	GL4	JA5	GP6	EM7	JK8
ON6	JO7	MD8	HQ9	GO10	LM1
QP9	MR10	QH1	IR2	IQ3	NO4
SR2	DB3	BL4	JS5	KS6	PQ7
HA5	EC6	FP7	AD8	RA9	RS10
JC8	HF9	AE10	BE1	BD2	HB3
LE1	KI2	IM3	CF4	EG5	JD6
NG4	NL5	OS6	HK7	HJ8	LF9
PI7	QO8	JG9	IL10	LN1	OI2
RK10	AR1	NK2	JM3	OQ4	QK5
AM3	JB4	PA5	NQ6	PR7	CP8
CO6	ME7	DH8	OR9	GB10	ER1
EQ9	PH10	EI1	PS2	ID3	SG4
GS2	SK3	FC4	DG5	KF6	HI7
OB5	CN6	RO7	EH8	MH9	BO10
QD8	FQ9	GQ10	FI1	OJ2	RF3
SF1	OA2	HL3	KN4	QL5	DE6
BH4	RD5	NE6	LO7	SB8	MG9
DJ7	BG8	CM9	MP10	CE1	AH2
FL10	EJ1	SD2	QA3	FH4	IP5
HN3	HM4	RI5	RB6	IK7	KR8
JP6	KP7	LP8	SC9	JL10	FG1
LR9	NS10	HE1	GJ2	MO3	MA4

FIGURE 5 - BALANCED INCOMPLETE FACTORIAL DESIGN

A TO S = SPEAKERS 1 TO 19

1 TO 10 = STARTING WORD OF THREE WORD PAIRS