Proceedings of The Institute of Acoustics

ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

R.L.Pratt

Royal Signals and Radar Establishment, Malvern

INTRODUCTION

Studies are currently being conducted at this establishment to examine the suitability of speech as an I/O channel for Army Battlefield Data Communications systems. This paper presents the results of a preliminary experiment designed to quantify the intelligibility of 7 commercially available text-to-speech synthesis systems. Speech intelligibility tests were performed using the Diagnostic Rhyme Test (DRT), and supplementary data obtained using semantic differential scaling techniques.

INTELLIGIBILITY TESTING

The systems tested were DECTALK, PROSE 2000, CALL TEXT, INFOVOX, COMPUTER CONCEPTS (BBC MICRO), TEXAS INSTRUMENTS SPEECH SYSTEM, NAMAL, and a control "system" comprising six male subjects (Figure 1).

The design and development of the DRT is described by Voiers in reference [1], but briefly it is a two choice speech intelligibility test comprising 96 rhyming word pairs that differ by a single acoustic feature (or "attribute") in the initial consonant (Figure 2). There are six such attributes; voicing, nasality, sustention, sibilation, graveness and compactness (A1-A6). The implementation of the DRT at the Royal Signals and Radar Establishment's Acoustic Test Facility is outlined in reference [2].

In this experiment word lists were generated by the speech synthesisers and recorded for subsequent processing. The lists were then replayed to a listening panel whose task is to select the word they thought they heard from the appropriate word pair displayed on a VDU. Two such lists were generated for each system and these are referred to as replications (R1,R2). Since it is the phonetic-to-acoustic performance that is of primary interest here, changes to word spellings were permitted to avoid mispronounciations.

The DECTALK system can produce nine different "voices", three of which were selected; Perfect Paul, Beautiful Betty and Rough Rita. The remaining devices were tested in their default states and details of their revision numbers etc. are given in Figure 1.

Thus a total of ten different samples were evaluated, and will be referred to in this paper by the generic term voice (V1-V10). The three DECTALK voices were run a second time at the end of the initial trial to test for any learning effects (V11-V13).

The voices were tested under two conditions; clear (S1), and with speech-like broadband noise added to give a speech-to-noise ratio of 0 dB(A) (i.e. equal amounts of speech and noise, S2). Both speech and noise levels were measured using a British Telecom Speech Voltmeter type SV6 in the active speech level setting. The signals were A weighted for measurement purposes only using a B&K 2610 Measuring Amplifier in the SV6 filter circuit. Listening subjects were seated in a quiet environment and used Sony MDR-31 light-weight headsets.

RESULTS

The results of an analysis of variance are shown in Figure 3. The first table shows the magnitude of the variability for Attributes, Speech-to-noise ratios, Voices and combinations of all the variables. For the purposes of the analysis, attribute scores were summed over the two lists (replications), and individuals were treated as random effects.

By examining the magnitude of the entries in the column headed F-values it may be seen that the two main sources of variability are speech-to-noise ratio, follwed by voices. In a separate analysis replications were not found to be a main effect.

In order to test the significance of the rank ordering of the voices, a Newman-Keuls test must be performed. The results of such a test, which was conducted separately for the two speech-to-noise ratios, are given in Figure 4 and allow a number of observations to be made:-

a). Of the 6 "re-runs" (the 3 DECTALK voices at both speech-to-noise ratios V11-V13), none of the scores was significantly different on re-test from the original score, although they were in general higher indicating the presence of a learning effect.

b). In clear conditions there was no significant difference in DRT scores between the human talkers, Beautiful Betty, Call Text and Rough Rita.

c). In 0 dB(A) speech-to-noise ratio conditions the human talkers are now scored significantly higher than the best synthesised voice, Rough Rita, who in turn is scored higher than Beautiful Betty and Perfect Paul.

d). The smallest drop in score when changing from clear to 0

dB(A) speech-to-noise ratio occurs for the human talkers (15.8%) and the largest drop is for TI-Speech (48.3%).

e). The biggest change in position occurs for the CALL TEXT system (ranked fourth in clear conditions and eighth at 0 dB(A)).

Interactions are indicated as significant when variables do not act independently to produce a DRT score. For example, the effect of adding noise to the speech signal has an effect on the DRT score depending on the synthesis system, and this manifests itself as a Voice x Speech-to-noise interaction. The effects of interactions can best be viewed graphically (see Figure 5). Human speech being more robust, has the smallest drop in score when noise is added. If no interaction was present, the lines would all be parallel.

ACCEPTABILITY ASSESSMENT

Listening subjects also completed a questionnaire (Figure 6) after each Diagnostic Rhyme Test. The questionnaire requires them to rate subjectively the Intelligibility, Effort (required to comprehend), Naturalness and Pleasantness. At the start of the very first session it was explained that they were about to hear samples from the test material (both clear and 0 dB(A) conditions) which covered the complete performance range, and that the entire range of the scales should be used when completing the questionnaire. This material was also replayed before each the start of each session. No other instructions were given to subjects on assessment criteria to be used when completing the rating scales.

These assessments were carried out as a result of hearing the DRT material only. Since this comprises isolated, monosylabic words occurring at a rate of 1 word every 1.3 seconds, the effect of prosody is excluded. It is intended to conduct further subjective assessments using sentence material.

RESULTS

The results from the rating scales were correlated against each other in addition to the DRT score, and the results shown in Figure 5. The scales are generally highly correlated amongst themselves with the exception of the Naturalness scale. It is also interesting to note that the correlation between DRT score and effort (r=-0.90) is very slightly higher than the correlation between DRT score and intelligibility (r=-0.88) but this difference is only very minor.

CONCLUSIONS

This study is a preliminary attempt to develop an assessment methodology to quantify the performance of text-to-speech synthesis systems. The relative performance of the samples tested to date indicate a set of complex relationships between intelligibility and experimental variables. DECTALK (Rough Rita) has been found to give the highest DRT scores when the speech signal has been degraded by additive noise at source. In clear conditions, with DRT scores in the 90's, the discrimination of the test is not so powerful and other assessment techniques may be more appropriate.

Data obtained through the use of rating scales correlate strongly with DRT scores, although the results cannot be applied to sentence material.

ACKNOWLEDGEMENTS

The author would like to thank; Ferranti Computer Systems and Logica for their assistance in making the word lists for the Prose 2000/CALL TEXT and INFOVOX/TI-SPEECH systems respectively, Digital Equipment Corporation for the loan of DECTALK, Mr. R.N.V.Bailey for the preparation of the test material and Dr. A.Belyavin for his assistance in the statistical analysis of the results.

The tests were administered by the Institute of Sound and Vibration Research (Southampton University), under contract to DA/RADIO.

REFERENCES

[1] W.D.Voiers, 'Diagnostic evaluation of speech intelligibility', Benchmark papers in acoustics, Vol. 11, Speech intelligibility and speaker recognition (M. Hawley, ed.) Dowden, Hutinson and Ross, Stroudsburg, (1977).

[2] R.L.Pratt, The assessment of speech intelligibility at RSRE, Proceedings of the Institute of Acoustics Vol. 6 pt. 4 pp 439-443, 1984.

Copyright © Controller HMSO, London,1986

ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

EXPERIMENTAL VARIABLES

VOICES

V1 =Beautiful Betty (Run 1)
V2 =Prose 2000 v1.2
V3 =Call Text 5010 v3.1
V4 =Infovox (provisional British English)
V5 =Computer Concepts Speech Rom (requires BBC Micro)
V6 =Rough Rita (Run 1)
V7 =Texas Instruments TI-Speech (v2.0)
V8 =Perfect Paul (Run 1)
V9 =Namal Type & Talk Speech Computer SC10
V10=HUMAN (control)
V11=Beautiful Betty (Run 2)
V12=Rough Rita (Run 2)
V13=Perfect Paul (Run 2)

SPEECH-TO-NOISE RATIOS

S1=Clear (no added noise)
S2=0 dB(A) Speech-to-noise ratio

ATTRIBUTES

A1=Voicing
A2=Nasality
A3=Sustention
A4=Sibilation
A5=Graveness
A6=Compactness

Figure 1. Experimental variables. Perfect Paul, Beautiful Betty and Rough Rita are all produced by DECTALK (v2.0), which is not available commercially in Europe.

VOICING		NASALITY		SUSTENSION	
Voiced	Unvoiced	Nasal	Oral	Sustained	Interrupted
veal	feel	meat	beat	vee	bee
bean	pean	need	deed	sheet	cheat
gin	chin	mitt	bit	vill	bill
dint	tint	nip	dip	thick	tick
zoo	sue	moot	boot	foo	pooh
dune	tune	news	dues	shoes	choose
voal	foal	moan	bone	those	doze
goat	coat	note	dote	though	dough
zed	said	mend	bend	then	den
dense	tense	neck	deck	fence	pence
vast	fast	mad	bad	than	dan
gaff	calf	nab	dab	shad	chad
vault	fault	moss	boss	thong	tong
daunt	taunt	gnaw	daw	shaw	chaw
jock	chock	mom	bomb	von	bon
bond	pond	knock	dock	vox	box

SIBILATION		GRAVENESS		COMPACTNESS	
Sibilated	Unsibilated	Grave	Acute	Compact	Diffuse
zee	thee	weed	reed	yield	wield
cheep	keep	peak	teak	key	tea
jilt	gilt	bid	did	hit	fit
sing	thing	fin	thin	gill	dill
juice	goose	moon	noon	coop	poop
chew	coo	pool	tool	you	rue
joe	go	bowl	dole	ghost	boast
sole	thole	fore	thor	show	so
jest	guest	met	net	keg	peg
chair	care	pent	tent	yen	wren
jab	gab	bank	dank	gat	bat
sank	thank	fad	thad	shad	sag
jaws	gauze	fought	thought	yawl	wall
saw	thaw	bong	dong	caught	taught
jot	got	wad	rod	hop	fop
chop	cop	pot	tot	got	dot

Figure 2. The Diagnostic Rhyme Test Vocabulary (taken from reference 2). Some changes were made to ensure correct pronounciation.

ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

ANALYSIS OF VARIANCE FOR VARIABLE :DRT

FACTOR	IDENTIFIER	LEVELS
I	INDIVIDUAL	8
A	ATTRIBUTE	6
S	SPEECH-TO-NOISE	2
V	VOICES	13

SOURCE OF VAR.	SUMS OF SQUARES	D.O.F.	MEAN SQUARES	ERROR TERM	F VALUE	PROB	
I	10034.4	7	1433.49				
A	60881.8	5	12176.4	AI	27.514	0.0000	***
AI	15489.4	35	442.555				
S	296161.	1	296161.	SI	434.909	0.0000	***
SI	4766.80	7	680.972				
V	188041.	12	15670.0	VI	145.516	0.0000	***
VI	9045.60	84	107.686				
AS	17187.8	5	3437.56	ASI	26.851	0.0000	***
ASI	4480.86	35	128.025				
AV	64039.8	60	1067.33	AVI	13.158	0.0000	***
AVI	34068.3	420	81.1151				
SV	27035.9	12	2252.99	SVI	39.255	0.0000	***
SVI	4821.12	84	57.3943				
ASV	36927.1	60	615.452	ASVI	10.319	0.0000	***
ASVI	25050.6	420	59.6443				
TOTAL	798031.	1247					

Figure 3 Analysis of variance table.

CLEAR

VOICE	MEAN			
5	61.2104			
9	65.9625			
7	80.0292			I
2	81.2667			I
4	83.6083			I
8	87.5771	I		
13	88.8125	I	I	
12	90.9625	I	I	I
6	91.6750	I	I	I
3	91.8146	I	I	I
1	92.4021	I	I	I
11	94.0229		I	I
10	95.6458			I

0 dB(A)

VOICE	MEAN			
5	29.9563			I
7	31.7229			I
9	39.9188		I	
4	40.2479		I	
2	42.3875		I	
3	49.6854			
8	59.7125	I		
13	62.1875	I		
1	64.2000	I		
11	64.3375	I		
6	69.4854			I
12	70.7813			I
10	79.8417			

Figure 4 Newman-Keuls test for differences between voices. The difference in mean scores for voices joined by a line of I's is not statistically significant at the 5% level.

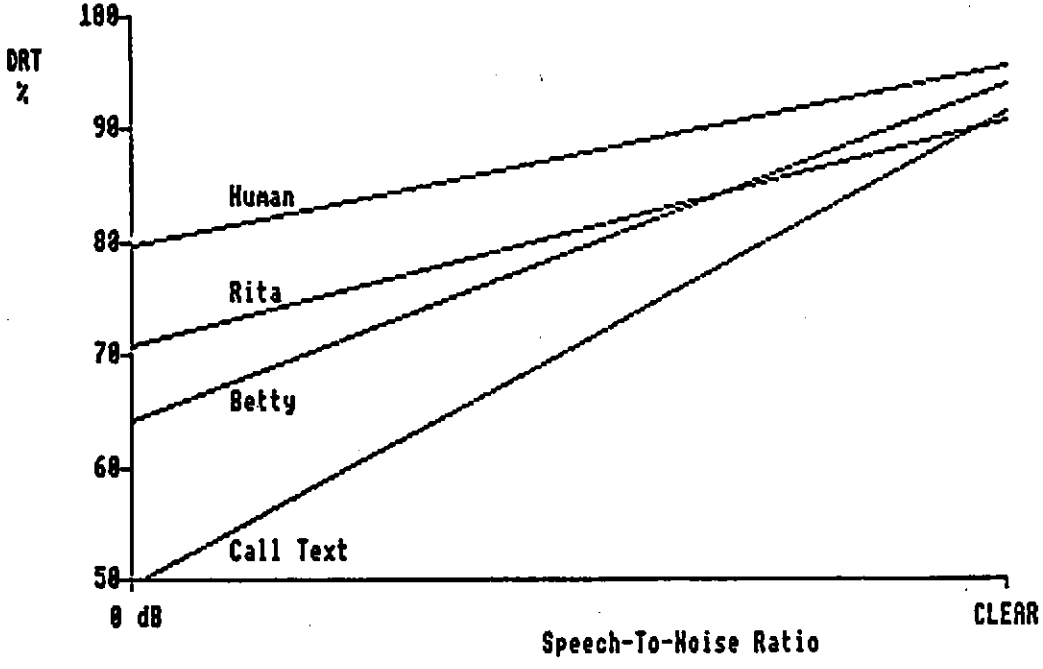

Figure 5 Example of Voice x Speech-to-Noise Ratio Interaction.

ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

INTELLIGIBILITY

Completely Intelligible	1	2	3	4	5	6	7	8	9	10	Totally Unintelligible

PLEASANTNESS

Very Pleasant	1	2	3	4	5	6	7	8	9	10	Very Unpleasant

NATURALNESS

Completely Natural	1	2	3	4	5	6	7	8	9	10	Totally Unnatural

EFFORT required to comprehend

No special Effort required	1	2	3	4	5	6	7	8	9	10	Extreme Effort required

Figure 6 The Semantic Rating Scale Questionnaire.

	DRT	INT	EFF	NAT	PLE
DRT SCORE	1.00				
INTELLIGIBILITY	-0.88	1.00			
EFFORT	-0.90	0.97	1.00		
NATURALNESS	-0.31	0.64	0.56	1.00	
PLEASANTNESS	-0.79	0.93	0.94	0.71	1.00

Figure 7 Correlation Matrix for DRT scores and subjective ratings.

ON THE INTELLIGIBILITY OF SYNTHETIC SPEECH

INTELLIGIBILITY

Completely Intelligible	1	2	3	4	5	6	7	8	9	10	Totally Unintelligible

PLEASANTNESS

Very Pleasant	1	2	3	4	5	6	7	8	9	10	Very Unpleasant

NATURALNESS

Completely Natural	1	2	3	4	5	6	7	8	9	10	Totally Unnatural

EFFORT required to comprehend

No special Effort required	1	2	3	4	5	6	7	8	9	10	Extreme Effort required

Figure 6 The Semantic Rating Scale Questionnaire.

	DRT	INT	EFF	NAT	PLE
DRT SCORE	1.00				
INTELLIGIBILITY	-0.88	1.00			
EFFORT	-0.90	0.97	1.00		
NATURALNESS	-0.31	0.64	0.56	1.00	
PLEASANTNESS	-0.79	0.93	0.94	0.71	1.00

Figure 7 Correlation Matrix for DRT scores and subjective ratings.