HUMAN SPEECH RECOGNITION PERFORMANCE ON THE 1994 CSR SPOKE 10 CORPUS

by

Will Ebel and Joe Picone
{ebel, picone}@ee.msstate.edu

Institute for Signal and Information Processing
Mississippi State University

PO Drawer EE
216 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-3649
Fax: 601-325-3149

GOALS

- Establish a reasonable target for machine performance

  Humans achieved a 1% word error rate

- Demonstrate that human performance on noisy data was high

  Word error rate does not degrade gracefully with SNR

- Calibrate performance as a function of SNR

  Human performance exceeds machines by at least 10dB

THE CSR'94 SPOKE 10 CORPUS

- Subset of the 5K-word Wall Street Journal Corpus (WSJ1)

- Total of 113x4 utterances subdivided as follows:

  Nominally 11 utterances/speaker

  10 speakers

  ☛ Four conditions: no noise, SNR = 22dB, 16dB, 10dB

- Additive Noise Characteristics

  Collected from a Nissan Maxima traveling at 62 m.p.h.

  Recorded using an omnidirectional microphone

  ☛ SNR per utterance varies; global signal level used

TESTING METHODOLOGY

- 12 subjects: 6 Male and 6 Female
  - ☛ English is first and primary language
    Normal hearing
    College-educated adults
    Computer literate
- Each subject transcribed 113 utterances
    Subjects arranged in three groups of four listeners
    Heard the same number of stimuli at each noise level
  - ☛ The entire Spoke 10 Corpus fully evaluated three times
    Each speaker hears a given prompt at only one noise level
- Evaluation setting
    Data entry simplified so as not to impact performance
    Subjects ONLY allowed to hear utterances in full
    No graphical tools, audio tools, or spectrograms allowed
    Subjects allowed to adjust volume but not spectral balance
    Ambient background noise reduced as much as possible
    Subjects participated in a training phase
    5000 word language model NOT imposed during evaluation
  - ☛ Subjects allowed to replay utterances as many times as desired and to modify any transcription at any time


ISIP
speech

SUBJECT DEMOGRAPHICS

Subject	Gender	Age (yrs)	Educ (yrs)	# Sess	Time (min.)	Errs (%)
01	M	32	8	1	150	1.6
02	M	19	0	1	105	2.7
03	F	24	6	1	180	2.2
04	F	39	8	1	165	0.6
05	M	18	0	1	150	2.4
06	F	35	5	1	195	2.0
07	F	33	7	1	180	1.9
08	M	29	8	2	330	0.5
09	F	40	8	2	210	1.8
10	F	28	8	1	150	2.5
11	M	17	0	1	135	4.5
12	M	25	4	1	150	2.1

Subjects are well educated adults providing a near upper bound on average human performance

Subject's transcription skills are very good

Note: Due to the "hidden agenda" of measuring the subjects' spelling abilities, the subjects have dictated that revealing their names with the scores will require a "black dot" security clearance!

COMBINED WORD ERROR RATES FOR ALL SUBJECTS

Evaluation Group	Vocabulary	
	Open	Closed
Average	2.1 (0.7)	1.0 (0.6)
Committee	1.2 (0.6)	0.5 (0.6)

Notes: Committee decisions were made on a word-by-word basis
Standard deviations are shown in parentheses

Overall human performance is at least an order of magnitude better than machine performance

OPEN-VOCABULARY WORD ERROR RATES

Listener	SNR				
	High	22 dB	16 dB	10 dB	Ave
Group 1:	1.8	2.0	2.0	1.6	1.9
l_01	1.6	3.2	2.0	1.7	1.9
l_02	2.7	1.8	2.8	4.3	2.8
l_03	2.2	0.8	1.8	0.9	1.5
l_04	0.6	2.6	2.0	0.7	1.3
Group 2:	1.8	2.2	1.9	2.0	2.0
l_05	2.4	4.0	2.2	2.8	2.7
l_06	2.0	1.6	0.9	2.6	1.7
l_07	1.9	1.0	2.2	1.3	1.6
l_08	0.5	2.7	3.0	1.4	1.8
Group 3:	2.5	2.2	2.5	2.7	2.5
l_09	1.8	1.7	1.2	2.9	1.7
l_10	2.5	2.0	3.3	4.0	2.8
l_11	4.5	1.9	3.9	2.2	3.2
l_12	2.1	3.1	2.2	2.2	2.3
All	2.0	2.1	2.1	2.1	**2.1**
Committee	1.0	1.4	1.2	1.2	**1.2**

Human performance is 1.2% for committee evaluations

Human performance does not vary with SNR

SPELLING-CORRECTED WORD ERROR RATES

Listener	SNR				
	High	22 dB	16 dB	10 dB	Ave
Group 1:	0.6	1.0	1.1	1.0	0.9
l_01	0.0	1.4	0.9	0.9	0.7
l_02	0.8	0.4	1.6	2.4	1.3
l_03	1.3	0.6	0.6	0.3	0.7
l_04	1.4	1.9	1.4	0.6	1.0
Group 2:	0.8	0.8	0.8	1.1	0.9
l_05	0.3	1.5	0.7	1.3	0.9
l_06	1.6	0.2	0.6	2.4	1.2
l_07	0.8	0.8	1.1	0.5	0.8
l_08	0.3	0.9	0.4	0.7	0.7
Group 3:	1.4	0.8	1.2	1.3	1.2
l_09	1.3	0.5	0.6	1.7	0.9
l_10	1.5	0.0	1.5	1.5	1.1
l_11	2.3	1.3	1.7	0.6	1.5
l_12	1.0	1.3	1.2	1.2	1.1
All	0.9	0.9	1.0	1.1	**1.0**
Committee	0.4	0.4	0.5	0.6	**0.5**

Human performance is 0.5% for committee evaluations

Human performance does not vary with SNR

SPELLING-CORRECTED AND NO-DUPLICATE PROMPTS

Listener	SNR				
	High	22 dB	16 dB	10 dB	Ave
Group 1:	0.9	0.7	0.7	0.9	0.8
l_01	0.0	0.7	0.3	1.3	0.4
l_02	1.1	0.4	1.2	3.1	1.2
l_03	1.6	0.7	1.0	0.0	0.8
l_04	0.9	1.1	0.7	0.7	0.7
Group 2:	1.0	0.9	0.7	1.0	0.9
l_05	0.3	1.3	1.3	1.4	1.0
l_06	2.7	0.0	1.2	2.4	1.3
l_07	0.5	0.6	1.4	0.6	0.7
l_08	0.7	1.3	0.0	0.5	0.7
Group 3:	1.2	0.7	1.0	1.5	1.1
l_09	1.3	0.0	0.7	2.4	0.9
l_10	1.9	0.0	1.2	2.2	1.2
l_11	1.6	1.0	1.2	0.4	1.1
l_12	0.5	1.6	1.2	1.2	1.1
All	1.0	0.8	0.8	1.1	0.9
Committee	0.6	0.3	0.4	0.7	0.5

Removing duplicate prompts does not significantly affect human performance

NO SIGNIFICANT CORRELATIONS BY SPEAKER

Speaker	SNR				
	High	22 dB	16 dB	10 dB	Ave
4t0	2.8	3.0	3.0	3.5	3.1
4t2	1.1	0.0	0.4	1.1	0.7
4t3	1.3	0.0	0.9	0.9	0.8
4t5	0.4	0.7	0.2	0.2	0.4
4ta	0.0	0.0	0.2	0.2	0.1
4tb	0.8	0.5	1.0	1.6	1.0
4tc	2.1	1.5	0.6	1.8	1.5
4te	0.0	1.0	0.3	0.3	0.4
4tg	1.6	0.9	1.6	1.8	1.5
4th	0.0	0.0	0.0	0.0	0.0

Note: Results are for the spelling-corrected no-duplicate prompt data

Human performance not correlated with SNR on a speaker by speaker basis

ERROR MODALITIES

Modality	Number of Errors
Inattention	25 (22%)
"the the"	12 (11%)
1 phone	37 (33%)
2 phones	34 (30%)
3 phones	3 (3%)
4 phones	0 (0%)
5 phones	1 (1%)

About 1/3 of the errors resulted from inattention and the "the the" anomaly

Nearly all the "valid" transcription errors were 1 and 2 phones long

A LIST OF ALL ERRORS FOR COMMITTEE TRANSCRIPTIONS

ID	Transcriptions: (R) Denotes Reference; (H) Denotes Human Hypothesis	High	22 dB	16 dB	10 dB
4T0C0304	(R) the **INDEX HAS** averaged fifty four %percent... (H) the **INDEXES ***** averaged fifty four %percent...	x	x	x	x
4T0C0305	(R) an a. t. and t. spokesman said the **THE** company's attorneys... (H) an a. t. and t. spokesman said the ******* company's attorneys...	x	x	x	x
4T2C0307	(R) directors also **APPROVED** an increase in the quarterly dividend... (H) directors also **PROVED** an increase in the quarterly dividend...	x	x	x	x
4T0C0308	(R) until a. t. and t.'s attorneys **FINISH** their review... (H) until a. t. and t.'s attorneys **FINISHED** their review...	x	x		x
4THC0301	(R) ...a number of parties have shown an interest in **INQUIRING** the unit... (H) ...a number of parties have shown an interest in **ACQUIRING** the unit...		x	x	x
4THC0306	(R) odyssey **PARTNERS** said it holds a five .point eight %percent stake... (H) odyssey **PARTNER** said it holds a five .point eight %percent stake...		x	x	
4T2C0303	(R) ...the quarter exceeded one dollar a share ***** union federal president... (H) ...the quarter exceeded one dollar a share **A** union federal president...	x			x
4TCC0301	(R) ...shareholder approval for the **PLAN** at its annual meeting... (H) ...shareholder approval for the **PLANT** at its annual meeting...	x			
4TGC0306	(R) we **CAN** compete (H) we **CAN'T** compete			x	x
4TCC030A	(R) ...all the economic indicators are solid **"QUOTE** and he attributed ... (H) ...all the economic indicators are solid **QUOTA** and he attributed ...	x			
4TCC0301	(R) ...and nuclear technology concern **SAID** it would seek shareholder... (H) ...and nuclear technology concern **SAYS** it would seek shareholder...			x	
4T2C0304	(R) the fully diluted figure reflects **A** forty .point three million... (H) the fully diluted figure reflects **THE** forty .point three million...				x
4TBC0305	(R) but he says he's cut back holdings **OF** public money managers (H) but he says he's cut back holdings **IN** public money managers				x
4TBC0309	(R) ...being asked to participate in the swap **AND** general electric credit (H) ...being asked to participate in the swap **IN** general electric credit				x

SUMMARY

- Human performance is high (average of 1% word error rate)

  Human performance is at least one order of magnitude better than machines

- No clear relationship between word error rate and SNR is evident which suggests:

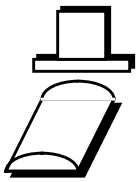
  Word error rate does not degrade gracefully with SNR
  (A sharp performance threshold most likely exists)

  Human performance exceeds machines by at least 10 dB

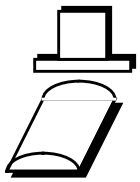
ACKNOWLEDGEMENTS

"We are indebted to those who sacrificed their lives for the advancement of speech science."

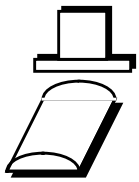
Sean Lauderdale



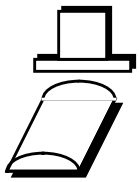
Stephanie Skinner



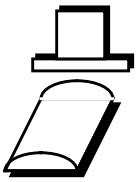
Mary Ann Picone



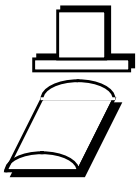
William Ebel



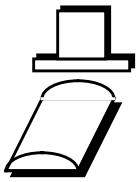
Daniel Williams



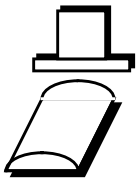
Jane Moorhead



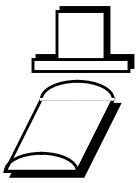
Rhonda Vickery



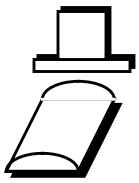
David Tannenbaum



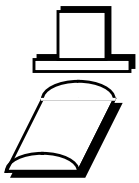
Debra Hicks



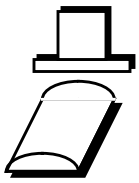
Regina Halpin



Richard Anton



Berry McCormick



- We promised our subjects they would become famous if they participated. The next time you meet one of these people on the street, ask them for their autograph!