

HUMAN SPEECH RECOGNITION PERFORMANCE ON THE 1995 CSR HUB-3 CORPUS

by

N. Deshmukh, A. Ganapathiraju, R. Duncan, and J. Picone

{deshmukh, ganapath, picone}@isip.msstate.edu

URL: <http://www.isip.msstate.edu>

Institute for Signal and Information Processing
Mississippi State University

ABSTRACT

Characterizing the differences between machine and human speech recognition performance continues to be a vital and important activity in speech research. While performance on limited vocabularies seems to be well understood, performance on expansive tasks such as those represented in Hub-3 is a more controversial issue. In this study we present benchmarks for fifteen listeners measured across four microphone conditions on data that involved more complex transcription challenges (e.g. surnames).

The error rates on the Hub-3 corpus were quite low — a 0.5% overall word error rate for a committee decision (ranging from 0.3% for the Audio Technica condenser to 0.8% for the Radio Shack electret). This is comparable to the results obtained on the CSR'94 corpus and is an order of magnitude better than the best machine performance on Hub-3. Most of the errors were due to inattention, supporting our perception that this year's task was more taxing on our listeners compared to last year's evaluation. In spite of these observations, human performance on Hub-3 was marginally better than that on the CSR'94 Spoke 10 corpus.



WHAT DID WE LEARN LAST YEAR?

- The CSR'94 Spoke 10 Corpus

Nominally 11 utterances/speaker, 10 speakers

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

- 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)

- Human performance was high and at least one order of magnitude better than machines
- No clear relationship between word error rate and SNR:

Listener	SNR				
	High	22 dB	16 dB	10 dB	Ave
Group 1:	0.6	1.0	1.1	1.0	0.9
Group 2:	0.8	0.8	0.8	1.1	0.9
Group 3:	1.4	0.8	1.0	1.3	1.2
All	0.9	0.9	1.0	1.1	1.0
Committee	0.4	0.4	0.5	0.6	0.5

- Human performance exceeded machines by at least 10 dB



CSR'95 HUB-3 CORPUS

Text Source: The North American Business (NAB) News

20 groups (articles) of 15 sentences each

Each article of 15 sentences was a contiguous excerpt drawn from a different article appearing in the August 1995 issue of a source

A judgement was made regarding the effort required to correctly read the article

Multiple Microphone Conditions

The data was balanced for sex (10 males/10 females)

Each group of 15 sentences was read by a different speaker and recorded simultaneously using two microphones

One of the microphones was fixed for all speakers — (a Sennheiser HMD-410 microphone (mic s))

Three alternates microphones:

mic_b: a Shure SM58 boom-mounted mic

mic_f: an Audio Technica AT851a Micro Cardioid Cond.

mic_g: Radio Shack 33-1060 Omni Elect. mic (desk stand)

Total of 600 utterances:

20 speakers x 15 utt./speaker x 2 mics/utt.



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

LISTENER ASSIGNMENTS

- ❑ Major Constraints:
 - 120 utterances per listener (time)
 - Minimize total number of listeners (volunteers)
 - Support a committee decision for all utterances (inattention)
- ❑ 15 listeners x 120 utterances gives desired coverage (900 utt. for mic_s; 900 utt. for alternate mics):

List #	Speaker index			
	Sennheiser mic	Mic b	Mic f	Mic g
1	711, 713, 715, 717	710, 716	712	718
2	710, 712, 714, 716	71j	719, 71d	711
3	719, 71b, 71d, 71f	71c	71i	713, 715
4	718, 71a, 71c, 71e	71g, 71h	714	71b
5	71g, 71h, 71i, 71j	717	71e, 71f	71a
6	71i, 71g, 71e, 71c	710	712	711, 713
7	71j, 71h, 71f, 71d	716, 717	714	715
8	71a, 718, 716, 714	71g	719, 71d	71b
9	71b, 719, 717, 715	71c, 71j	71e	718
10	713, 712, 711, 710	71h	71f, 71i	71a
11	710, 715, 71a, 71f	71g	719	713, 71b
12	711, 716, 71b, 71g	710	71e, 71i	718
13	712, 717, 71c, 71h	71j, 716	71d	71a
14	713, 718, 71d, 71i	71h	714, 71f	711
15	714, 719, 71e, 71j	717, 71c	712	715



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

AN OVEVIEW OF THE LISTENER POPULATION

Listener	Sex	Age (yrs)	Time (min)	Errors (%)	
				v1.0	v1.1
01	M	19	225	2.1	2.0
02	F	23	135	1.7	1.5
03	F	22	190	4.0	3.3
04	F	23	215	4.1	3.0
05	M	22	180	1.8	1.7
06	F	24	110	1.5	1.4
07	M	23	160	1.7	0.9
08	M	24	157	2.0	1.4
09	M	23	225	2.1	2.0
10	M	31	115	2.0	1.4
11	M	23	270	3.7	3.2
12	F	26	210	4.1	4.0
13	F	21	195	2.4	2.2
14	M	28	150	3.6	3.5
15	F	40	305	2.1	1.2
Average		25	190	2.7	2.2

Note: Listeners 1 and 15 participated in the CSR'94 experiment. All other listeners were first-time participants.



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

COMBINED WORD ERROR RATES FOR ALL SUBJECTS

Vocab type	Listener	Error (%) on corpus	
		Spoke 10	Hub - 3
Open	Overall	2.1	2.2
	Committee	1.2	0.8
	Listener 01	2.7	2.0
	Listener 15	1.3	1.2
Augmt.	Overall	1.0	2.1
	Committee	0.5	0.4
	Listener 01	1.0	2.0
	Listener 15	0.7	1.1

Notes:

Results comparable with CSR'94 Spoke 10

Larger difference between overall and committee results for Hub-3 augmented vocabulary condition (an indication of the difficulty of the task)



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

DETAILED OPEN-VOCABULARY RESULTS

Listeners & groups	Word error (%) on microphone conditions				
	Mic s	Mic b	Mic f	Mic g	All
Group 1	2.1	2.3	2.0	3.0	2.3
Listener 01	2.3	1.0	4.2	0.3	2.0
Listener 02	1.3	2.1	1.2	2.3	1.5
Listener 03	3.3	4.7	3.4	2.2	3.3
Listener 04	2.5	1.1	2.3	9.9	3.0
Listener 05	1.6	5.0	0.4	0.6	1.7
Group 2	1.4	1.3	0.9	2.2	1.4
Listener 06	1.6	0.0	0.6	1.9	1.4
Listener 07	0.4	2.2	1.0	0.0	0.9
Listener 08	1.1	0.8	0.4	5.4	1.4
Listener 09	2.3	1.9	0.3	2.4	2.0
Listener 10	1.5	0.6	2.0	1.3	1.4
Group 3	2.5	2.1	2.8	4.7	2.8
Listener 11	1.2	3.0	0.4	8.4	3.2
Listener 12	4.0	1.5	4.8	5.1	4.0
Listener 13	3.0	1.0	1.7	1.3	2.2
Listener 14	3.6	2.5	3.3	4.6	3.5
Listener 15	0.4	2.6	1.5	0.4	1.2
Overall	2.0	1.9	1.9	3.3	2.2
Committee	0.8	1.0	0.2	1.5	0.8
Aug. Com.	0.3	0.6	0.1	0.8	0.4



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

PERFORMANCE AS A FUNCTION OF SPEAKER

Speaker	Word error (%) on microphone conditions					Speech Rate (words/min.)
	Mic s	Mic b	Mic f	Mic g	All	
710	1.0	1.0	-	-	1.0	192
711	2.3	-	-	3.3	2.8	184
712	2.8	-	2.5	-	2.7	217
713	1.4	-	-	2.8	2.1	205
714	0.4	-	2.0	-	1.2	158
715	1.4	-	-	0.7	1.1	212
716	2.0	0.8	-	-	1.4	176
717	5.4	4.0	-	-	4.7	181
718	3.0	-	-	2.3	2.6	227
719	1.0	-	1.2	-	1.1	162
71a	0.6	-	-	1.2	0.9	187
71b	5.7	-	-	8.7	7.2	232
71c	3.7	2.8	-	-	3.2	172
71d	1.5	-	0.7	-	1.1	210
71e	1.6	-	1.5	-	1.5	197
71f	0.2	-	1.7	-	0.9	165
71g	1.8	1.6	-	-	1.7	240
71h	0.5	1.5	-	-	1.0	188
71i	2.5	-	4.1	-	3.3	174
71j	1.4	1.3	-	-	1.3	170



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

WORD ERROR CATEGORIES

	Mic s	Mic b	Mic f	Mic g	Total
Proper Nouns	25	8	1	10	44 (45.8%)
Listener	5	1	-	4	10 (10.4%)
Articulation	1	1	1	1	4 (4.2%)
Inattention	14	10	1	9	34 (35.4%)
Signal	-	2	-	2	4 (4.2%)
Total	45 (46.9%)	22 (22.9%)	3 (3.1%)	26 (27.1%)	96 (100)



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

AUGMENTED COMMITTEE DECISION RESULTS

Listener	Word error (%) on microphone conditions (noise levels)				
	Mic s (38 dBA)	Mic b (21 dBA)	Mic f (21 dBA)	Mic g (19 dBA)	Over-all
Humans	0.3	0.6	0.1	0.8	0.4
Machine	6.6	5.1	5.4	9.9	6.7

AUGMENTED WORD ERROR CATEGORIES

	Mic s	Mic b	Mic f	Mic g	Total
Listener	5	1	-	4	10 (19.2%)
Articulation	1	1	1	1	4 (7.7%)
Inattention	14	10	1	9	34 (65.4%)
Signal	-	2	-	2	4 (7.7%)
Total	20 (38.5%)	14 (26.9%)	2 (3.8%)	16 (30.8%)	52 (100%)



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

TYPICAL ERRORS

Listener Errors:

id: (717c0202/mic_ss)

REF: ... compelling them to EMIGRATE to the u. s.

HYP: ... compelling them to IMMIGRATE to the u. s.

id: (71jc020a/l02_mic_ab)

REF: it seems to be the fraud DU jour said allen hile...

HYP: it seems to be the fraud DE jour said allen hile...

Signal Errors:

id: (713c020a/mic_ag)

REF: AND this has all happened...

HYP: * this has all happened...

Inattention:

id: (712c020a/mic_af)

REF: given the demand it is expected **** i. b. m. will...

HYP: given the demand it is expected THAT i. b. m. will...

Articulation:

id: (71gc0203/mic_ab)

REF: ...network shares TRADED in the low sixty dollar...

HYP: ...network shares TREATED in the low sixty dollar...

Other Interesting Errors:

id: (711c0206/mic_ag)

REF: ...more of a speculation said GIAN CAMUZZI senior...

HYP: ...more of a speculation said JION CAMUTSIE senior...

id: (711c0207/mic_ag)

REF: ...and goldman SACHS AND company...

HYP: ...and goldman * SAXON company...

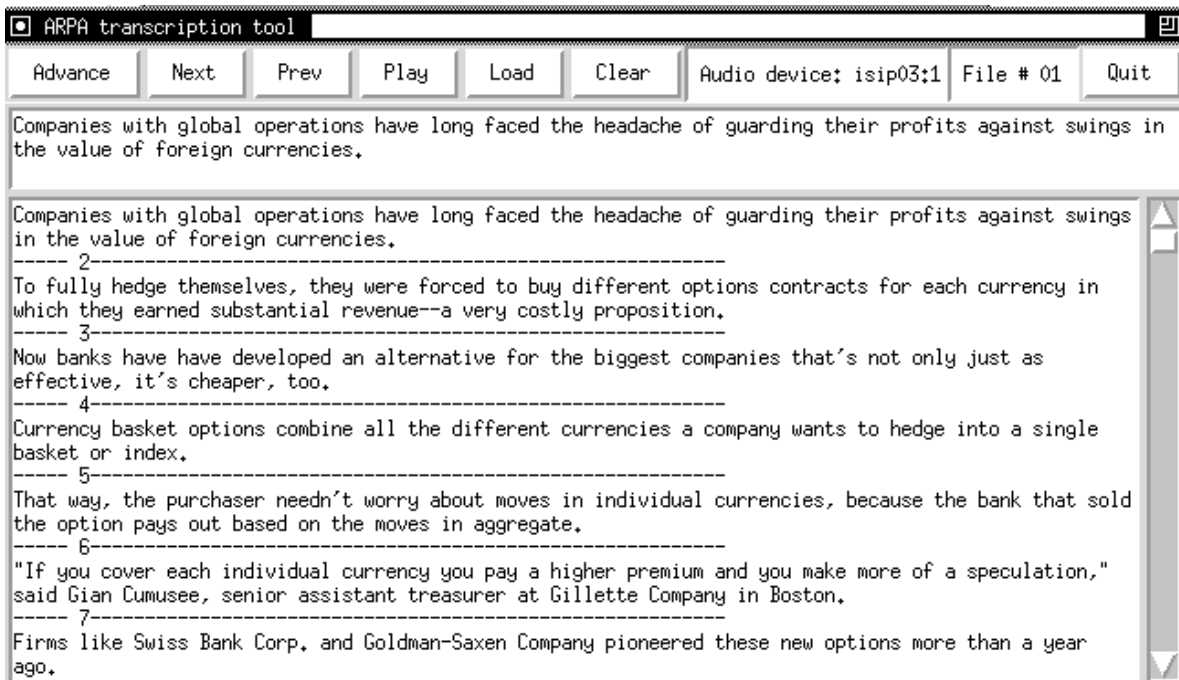
Note: audio tape contains mic_ss followed by the actual mic for the condition in which the error occurred.



INSTITUTE FOR SIGNAL AND INFORMATION PROCESSING

WHAT ABOUT NEXT YEAR?

- Our subjects have requested that you do not make the task any harder! (my wife is tired of cooking brownies...)
- More control over listening (windowing):



ARPA transcription tool

Advance Next Prev Play Load Clear Audio device: isip03:1 File # 01 Quit

Companies with global operations have long faced the headache of guarding their profits against swings in the value of foreign currencies.

Companies with global operations have long faced the headache of guarding their profits against swings in the value of foreign currencies.
----- 2-----
To fully hedge themselves, they were forced to buy different options contracts for each currency in which they earned substantial revenue--a very costly proposition.
----- 3-----
Now banks have have developed an alternative for the biggest companies that's not only just as effective, it's cheaper, too.
----- 4-----
Currency basket options combine all the different currencies a company wants to hedge into a single basket or index.
----- 5-----
That way, the purchaser needn't worry about moves in individual currencies, because the bank that sold the option pays out based on the moves in aggregate.
----- 6-----
"If you cover each individual currency you pay a higher premium and you make more of a speculation," said Gian Cumusee, senior assistant treasurer at Gillette Company in Boston.
----- 7-----
Firms like Swiss Bank Corp. and Goldman-Saxen Company pioneered these new options more than a year ago.



SUMMARY

- The Hub-3 evaluation yielded results consistent with those obtained for Spoke 10:

Performance of common listeners was comparable

Human performance is high (average of 0.4% word error rate)

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

- Human performance was also fairly consistent across all the microphones
- Context did not play a significant role
- Listener inattention is becoming an increasingly significant problem

