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

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

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762
{deshmukh, ganapath, duncan, picone}@isip.msstate.edu

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.

1. INTRODUCTION TO HUB-3 CORPUS

It is a well-accepted fact that human performance significantly exceeds machine performance on a wide range of speech recognition tasks [1]. In the CSR'94 evaluations, we established that human recognition performance is unaffected even at signal-to-noise ratios (SNRs) as low as 10dB [1]. In contrast, machine performance was shown to degrade significantly at such low SNRs. SNR only tells part of the story, however. To achieve a more meaningful benchmark for comparison of machine performance it is important to calibrate the effect of other acoustic correlates such as microphone type and microphone placement. Another equally important dimension of the recognition task, particularly relevant to NAB and WSJ-type corpora, that we address in this study is out-of-vocabulary words. The Hub-3 corpus was developed to focus on the evaluation of these aspects of the recognition problem.

The CSR'95 Hub-3 [2] evaluation test data consists of 20 groups of 15 sentences each. Each group of 15 sentences was a contiguous excerpt drawn from a different article appearing in the August 1995 issue of a North American Business (NAB) news source (one of WSJ/DJIS, NYT, LAT, WP and REU/RNAB). The articles were selected at random and examined for content. A judgement based on several explicit criteria was made regarding the readability of the articles. If an article was judged too difficult to read, another was picked in its place. (Despite this careful screening, several data files did contain anomalous pronunciations.)

The data was balanced for sex (10 males/10 females) where each of the twenty subjects was a native speaker of American English. 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. For this reference condition, the standard ARPA CSR 'close-talking' microphone — a Sennheiser HMD-410 microphone (mic s) — was employed. The second condition consisted of a microphone selected from one of three alternates — a Shure SM58 boom-mounted mic (mic b), an Audio Technica AT851a Micro Cardioid Condenser Boundary mic (mic f), and a Radio Shack 33-1060 Omni Electret mic in a slip-on desk stand (mic g). All speech was recorded in a somewhat reverberant room with an ambient noise level of 72 dB SPL (linear) and 54 dB (A-weighted). Thus the corpus consists of a total of 600 utterances — 20 speakers x 15 utt./speaker x 2 mics/utt.

2. EXPERIMENTAL DESIGN

A successful human performance benchmark involving volunteer listeners must minimize training phenomena. Based on last year's experiences [1], we expected some listener adaptation to speaker, microphone, and material to occur. We primarily sought to calibrate performance as a function of the microphone characteristics and to limit performance variations due to such second-order effects. We developed a test methodology that minimizes the number of utterances each listener is required to evaluate, and also reduces the number of listeners required, without significantly affecting the overall results.

2.1. Experimental Setup

Our approach [1], required each listener to:

- listen to 120 utterances — 60 utterances (4 sets) from the Sennheiser microphone and 60 utterances (1 to 2 sets each) from the other three microphones;
- iterate over the entire data set as much as desired — no constraints were placed on the order of transcription, correction, or the amount of time spent on the task;

- listen to the data using Sony MDR-V50 headphones and a high quality 16-bit audio system (a Townshend DAT-Link+ and a Sony 60ES DAT player);
- alter only the volume level of the reproduction.

They had no *a priori* knowledge regarding the microphone conditions or the speaker. They were instructed that utterances were presented as a group of 15 sentences (this was fairly obvious from listening to the data). Use of any grammar or word processing tools, or any means of viewing waveforms or spectrograms of the utterances was prohibited. The subjects could only refer to the 64,000 word Hub-3 lexicon during transcription.

User Interface: A significant issue in conducting these evaluations involved the mechanics of transcription entry. Since the typical listener is not used to transcribing what they hear, the chance of artifacts appearing in the data is great. We minimized this risk by selecting knowledgeable subjects and using normal orthography for transcription. The data was then converted off-line, under the guidance of the researchers, to the proper evaluation format [3]. New to this year's evaluation was a tcl-based graphical user interface, shown in Figure 1, to facilitate transcription process.

2.2. Selection Of Listeners

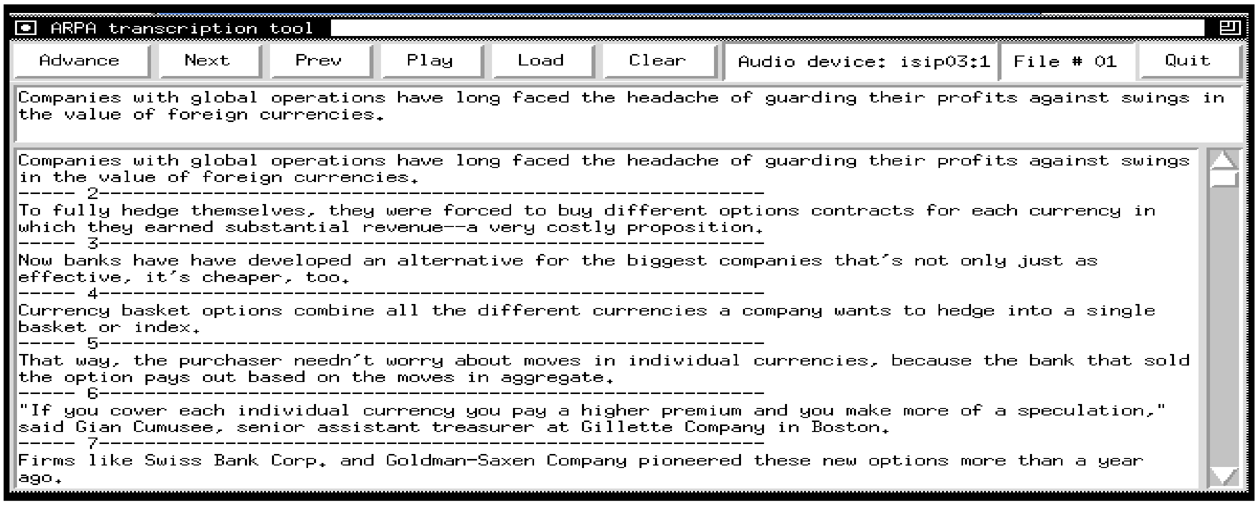
An overview of the listener population is given in Table 1. Listeners with the following characteristics were selected:

- 13 of the 15 subjects were native-born American citizens for whom English is the first and primary language. Of these, 2 had also participated in last year's evaluations;
- 2 subjects were non-native speakers of English who have been residents of the USA for more than two years;
- all subjects were college-educated adults with at least basic proficiency in the use of computers (Unix-based systems);
- the subjects were balanced for sex (8 males and 7 females).

Non-native speakers were introduced into this year's evaluation with the expectation that satisfactory performance would greatly increase the pool of listeners from which we can draw for future experiments. Two listeners who participated in last year's experiment were included in this year's evaluation in order that we could better calibrate performance as a function of corpus.

Listener	Sex	Age (yrs)	Time (min)	Errors (%)	
				v1.0	v1.1
01	M	19	225	2.2	2.0
02	F	23	135	1.6	1.5
03	F	22	190	4.0	3.3
04	F	23	215	4.2	3.0
05	M	22	180	1.7	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	2.3	2.0
10	M	31	115	2.0	1.4
11	M	23	270	3.9	3.2
12	F	26	210	4.0	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	2.6	2.2

Table 1: An overview of listener demographics and nominal performance. v1.0 corresponds to results for the raw data, while v1.1 represents error rates after spelling corrections.



```
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.
```

Figure 1: Transcription tool for the Hub-3 human benchmark evaluations

Stimuli Preparation For Each Listener: The distribution of utterances among the subjects is documented in Table 2. The listeners were split into three groups (as shown in Table 2) such that each group would hear each of the 600 utterances exactly once, in order that a three-way committee decision could be constructed. Some constraints on this randomization are as follows:

- each listener processes 120 utterances;
- each listener hears only one version of any sentence;
- each listener hears 60 utterances recorded on the close-talking microphone, 30 from one of the alternate microphones and 15 each from the other two;
- microphone conditions and speakers are presented in a fairly random order throughout the session.

Since each group covers the entire test set, our results can be collated to form four separate benchmarks for Hub-3 that can be compared directly to machine performance.

3. EVALUATION RESULTS

Most listeners chose to transcribe the data in a single session, the remainder required a second session. We denote this output version v1.0. After reviewing these transcriptions, it became clear that fatigue had caused many listeners to delete large sections of at least one to two utterances. Hence, we invited most of the listeners to review their transcriptions a second time. These review sessions typically occurred one week or more after the first session. We denote these transcriptions version v1.1. Not all listeners were available for a second pass (this highlights the importance of the committee decision result).

The evaluation was conducted as an open-vocabulary test. The listeners were given access to the 64K vocabulary [4], and were allowed to look up words in this file using a simple search tool (emacs). Unfortunately, most listeners were not as diligent as we would have liked, and consequently entered misspelled words (another example of motivational issues raised later). The overall performance of all the listeners is summarized in Table 3. The accuracy of the reference transcriptions this year was higher than last year. Nevertheless, several minor corrections had to be applied to the v1.1 data. For example, the pronunciation of the word “shoe” contained a word-final “d” as in “white SHOED law firm.” This was corrected to “shoe” to match the reference transcription.

As a postprocessing step, the transcriptions were converted into an *augmented-vocabulary* set (see Tables 3 and 5) by applying spelling corrections. The authors reviewed the transcriptions and corrected misspellings. Surnames whose spellings did not match well-known surnames and other such proper-noun terms were replaced with the correct spelling (as defined by the reference transcription). For example, when “Camutzi” appeared in a transcription (closely reflecting the pronunciation), it was replaced with the correct spelling, “Camuzzi,” in the hypothesis string. Note that the test set contained only 52 words (out of 1886 unique words in reference texts) that were not in the 64K vocabulary. We refer to the augmented-vocabulary set as version v2.0.

The committee decisions in the table reflect the transcriptions generated by pooling data for each utterance from all three listeners on a word-by-word basis and applying a majority vote. We expected this to eliminate most errors due to inattention by the subjects (unintentional errors) and to resolve some of the ambiguities resulting from particular listeners’ unfamiliarity with the topic. Listeners’ familiarity with the specific domains and their specialized terms played a significant role in the overall performance. The error rate dropped considerably when spelling corrections (mostly for proper nouns) were applied (Table 3).

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

Table 2: Distribution of speaker-microphone configurations for various listeners and group assignments

Evaluation group	Error (%) on vocabulary type	
	Open	Augmented
Overall	2.2	1.6
Committee	0.8	0.4

Table 3: An overview of human performance on the Hub-3 corpus. The committee decisions corresponds to a majority vote across three candidate transcriptions for each utterance.

3.1. The Open Vocabulary Evaluation

A detailed analysis of the open-vocabulary evaluation (v1.1) is presented in Table 4. Recall that each group consists of 5 listeners and hears each utterance once for all microphone conditions. Thus groups 1, 2 and 3 evaluated the same data and their results are directly comparable. Since an error of 0.05% is the equivalent of a

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**

Table 4: Detailed overview of the open-vocabulary evaluation

Listeners & groups	Word error (%) on microphone conditions				
	Mic s	Mic b	Mic f	Mic g	All
Group 1	**1.7**	**1.7**	**1.4**	**1.9**	**1.7**
Listener 01	1.2	0.3	2.4	0.3	1.0
Listener 02	0.8	1.7	1.0	1.6	1.1
Listener 03	3.0	4.5	2.6	1.3	2.8
Listener 04	2.4	1.1	1.9	5.7	2.4
Listener 05	1.3	2.3	0.0	0.6	1.1
Group 2	**1.0**	**1.0**	**0.8**	**1.5**	**1.0**
Listener 06	1.4	0.0	0.3	0.9	1.0
Listener 07	0.4	0.9	1.0	0.0	0.5
Listener 08	1.1	0.8	0.4	3.2	1.2
Listener 09	0.8	1.9	0.3	2.4	1.2
Listener 10	1.1	0.6	1.8	1.3	1.2
Group 3	**1.9**	**1.5**	**2.6**	**3.3**	**2.1**
Listener 11	0.9	2.3	0.4	5.7	2.2
Listener 12	2.8	1.2	4.3	5.1	3.2
Listener 13	2.0	0.9	1.7	1.3	1.6
Listener 14	3.2	2.5	3.3	2.3	3.0
Listener 15	0.3	1.3	1.5	0.0	0.7
Overall	**1.5**	**1.4**	**1.6**	**2.2**	**1.6**
Committee	**0.3**	**0.6**	**0.1**	**0.8**	**0.4**

Table 5: Results of the augmented-vocabulary evaluation

single word error for most conditions, for most of the results the differences are not statistically significant.

A look at the amount of agreement among listeners yields some interesting statistics. A group of three listeners who transcribed the same data (in different orders) disagreed on at least one word on 43% of the utterances. However, 90% of these could be resolved by a majority vote, and the remaining 10% of these differences generally pertained to proper nouns. In cases where all three listeners had a disagreement (typically only 2 to 3 of the 600 utterances) the transcription closest to the truth was selected.

3.2. The Augmented-Vocabulary Evaluation

The results for the augmented-vocabulary evaluation are given in Table 5. Spelling corrections resulted in a significant decrease in the error rate, reflecting the importance of the problem of proper-noun recognition. The open-vocabulary set had transcription errors in 28% of the sentences. In the augmented-vocabulary set it dropped to 22%. The committee reported sentence errors of 13% and 7% respectively on the two sets. Sentence errors were not a function of the length of the sentence, thus demonstrating the power of context in recognition. The most common error modalities were equally distributed amongst all standard categories.

The committee disagreed on only 25% of the utterances, most of which were resolved by a simple majority rule. Cases where there was no majority were typically inattention errors and were resolved arbitrarily by selecting the closest match to the reference transcription (again this only happened on a handful of utterances). The word error rate dropped 80% from the group case to the committee decision, resulting in 52 errors for 11,994 words. We believe the difference in the committee decision and the individual transcriptions would shrink if we were to use highly-trained transcribers. A large percentage of the errors appear to be motivated by inattention rather than by unfamiliarity with the subject matter.

Speaker	Word error (%) on microphone conditions				
	Mic s	Mic b	Mic f	Mic g	All
710	1.0	1.0	-	-	1.0
711	2.3	-	-	3.3	2.8
712	2.8	-	2.5	-	2.7
713	1.4	-	-	2.8	2.1
714	0.4	-	2.0	-	1.2
715	1.4	-	-	0.7	1.1
716	2.0	0.8	-	-	1.4
717	5.4	4.0	-	-	4.7
718	3.0	-	-	2.3	2.6
719	1.0	-	1.2	-	1.1
71a	0.6	-	-	1.2	0.9
71b	5.7	-	-	8.7	7.2
71c	3.7	2.8	-	-	3.2
71d	1.5	-	0.7	-	1.1
71e	1.6	-	1.5	-	1.5
71f	0.2	-	1.7	-	0.9
71g	1.8	1.6	-	-	1.7
71h	0.5	1.5	-	-	1.0
71i	2.5	-	4.1	-	3.3
71j	1.4	1.3	-	-	1.3

Table 6: Recognition performance as a function of speaker, averaged across all listeners

Given that we did not use trained transcribers, and that the subjects were not given any incentive to perform well, it appears as though the drastic reduction in error rate from the groups to the committee is not surprising. Since a single word error of this type can often modify the error rate by at least 0.05%, such anomalous errors have a profound impact on our ability to perform detailed analyses.

3.3. Analysis Of Errors

No significant correlation was found between microphone conditions s, b, and f. We believe this is because the SNR levels of these three microphones were simply not low enough to affect transcription for most listeners (the approximate A-weighted SNR range was 19 dB for mic g, 21 dB for mic b and mic f and 38 dB for mic s). Had the SNR been lower, a greater sensitivity may have resulted. Microphone g, on the other hand, was relatively worse than the other three microphones. There was one combination of speaker and microphone — speaker 71b and microphone g, that was definitely much harder to transcribe than the other datasets. A summary of error-rates per speaker (across all listeners) is presented in Table 6.

Error modality	Number of errors	
Inattention errors	286	(36.7%)
Habitual errors	65	(8.3%)
Errors due to 1 phone	262	(33.6%)
Errors due to 2 phones	100	(12.8%)
Errors due to 3 phones	31	(4.0%)
Errors due to 4 phones	20	(2.6%)
Errors due to 5+ phones	16	(2.0%)

Table 7: Common error modalities across all listeners: we see that a large percentage of sentence errors were caused by inattention, habit and one/two phone confusions.

We found three major classes of errors: inattention, habitual errors (insertions/substitutions — especially of articles like ‘a’, ‘the’ etc.) and valid auditory-based transcription errors. We further subdivided the latter category into the number of phone errors that constituted each word error (see Table 7). As expected, a large proportion of sentence errors involved morpheme-sized units (typically a function word); and most word errors involved only one or two phonemes. A significant portion of such errors involved proper nouns.

Finally, a few error modalities from the committee results for the augmented-vocabulary committee test set are presented in Table 8. As expected, humans outperformed the best machine results [4] on this corpus by at least an order of magnitude (see Table 9).

Comparison with Spoke 10 results: A comparison of the results with Spoke 10 is presented in Table 10. Note that listeners 1 and 15 participated in both evaluations, and had comparable performance. The committee results were identical in the augmented vocabulary case. The degradation in average errors is probably due to the complexity of Hub-3 which contains terminology in specialized domains that many of the listeners may be unfamiliar with, thus increasing inattention.

Results with non-native listeners: Two of the subjects participating in this evaluation, listeners 07 and 13, were foreign nationals whose native or primary language is not English. Yet their performance was on par with the average if not better, both on the open and augmented vocabulary tests. Both of these listeners were known to have outstanding English skills.

4. SUMMARY

The Hub-3 evaluation yielded results consistent with those obtained for Spoke 10. Error rates for humans were low, machine performance on the same data is an order of magnitude worse.

Human performance was also fairly consistent across all the

Spk# (Utt#)	Transcriptions (R) denotes Reference, (H) denotes Human hypothesis	Error on utterance for			
		Mic s	Mic b	Mic g	Mic f
717 c0202	(R): ... compelling them to **EMIGRATE** to the u. s. (H): ... compelling them to **IMMIGRATE** to the u.s.	✓			
713 c0201	(R): **YOU** could learn a lot from a so called slacker (H): **WE** could learn a lot from a so called slacker			✓	
71b c0208	(R): **THE GROUP** together with the three new co-chairmen formed a... (H): **THEY GROUPED** together with the three new co-chairmen formed a...	✓			
712 c0208	(R): ... chips that store two hundred **AND** fifty six million bits of information (H): ... chips that store two hundred ******* fifty six million bits of information	✓			✓

Table 8: Characteristic error examples for committee evaluations on extended-vocabulary evaluations

Listener	Word error (%) on microphone conditions[1] (noise levels)				
	Mic s (38 dBA)	Mic b (21 dBA)	Mic f (21 dBA)	Mic g (19 dBA)	Over -all
Humans	0.4	0.5	0.3	0.8	0.5
Machine	6.6	8.1	10.3	23.9	10.0

Table 9: Comparison of human performance on Hub-3 with the best machine performance as a function of microphone (and SNR levels). The machine results are from the Cambridge HTK-based system [4].

Vocab type	Listener	Word error (%) on corpus	
		Spoke 10	Hub - 3
Open	Overall	2.1	2.2
	Committee	1.2	0.5
	Listener 01	2.7	2.1
	Listener 15	1.3	1.2
Augmt.	Overall	1.0	1.9
	Committee	0.5	0.5
	Listener 01	1.0	1.8
	Listener 15	0.7	0.7

Table 10: Comparison with the Spoke 10 human benchmark

microphones. It appears that humans can discern speech from noise quite well in the range of SNR levels used in this experiment and the artifacts introduced by the microphone. Context also plays an important role in separating the speaker from noise, though in spite of the availability of more contextual information in Hub-3 the recognition performance is comparable to Spoke 10.

Humans seem to perform consistently over time and over different domains. Performance of common listeners on Spoke 10 and Hub-3 corpora is comparable. The excellent performance of the foreign-national subjects demonstrates that non-native speakers can also participate in such evaluations, thus widening the pool of listeners at our disposal.

A large percentage of the errors in the extended-vocabulary case were due to listener inattention. By providing sufficient incentive to listeners to avoid errors due to attention lapses we can generate benchmarks for human recognition that are more meaningful from a machine-comparison perspective. It appears that the current corpora tax the limits of human performance on a “volunteer” basis.

5. ACKNOWLEDGEMENTS

First and foremost, we are grateful to our fifteen listeners who donated their time and energy to this study. Many others contributed to this deceptively simple project to insure a quality result. We thank George Doddington and Dave Pallett for creating the opportunity to conduct this experiment, for providing excellent insight into the problem, and for many valuable discussions about the interpretation of the results.

REFERENCES

1. Ebel, W.J. and Picone, J., “Human Speech Recognition Performance on the 1994 CSR SPOKE 10 Corpus”, in *Proceedings of the SLST Workshop,* pp. 53-59, Austin TX, January 1995.

2. Pallett, D., “CSR’95 Hub-3 Multi-Microphone Evaluation Test Data,” NIST Speech Disc R27-6.1, NIST, Room A216 Building 225 (Technology), Gaithersburg, MD 20899, October 26, 1995.

3. Pallett, D., “System Output Preparation and Scoring Protocols,” NIST, Room A216 Building 225 (Technology), Gaithersburg, MD 20899, October 4, 1994.

4. Pallett, D., et. al., “1995 Benchmark Tests for the ARPA Spoken Language Program,” in *Proceedings of the SLST Workshop*, Harriman NY, February 1996.

[1] Mic s — Sennheiser HMD-410 close-talking
Mic b — Shure SM-58 boom-mounted
Mic f — Audio Technica AT851a Micro Cardioid Condenser
Mic g — Radio Shack 33-1060 Omni Electret