HUMAN SPEECH RECOGNITION PERFORMANCE ON THE 1994 CSR SPOKE 10 CORPUS

W.J. Ebel and J. Picone

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, Mississippi 39762
{ebel, picone}@ee.msstate.edu

ABSTRACT

Recently, questions have been raised about the ability of humans to recognize and transcribe noisy speech data. To benchmark human performance, we conducted an experiment involving twelve human listeners and the CSR'94 Spoke 10 Corpus. The experiment was designed so that each listener heard the spoken response to a prompt only once, every utterance in the corpus was heard exactly three times, and subjects could revise any transcription at any time.

This experiment resulted in two important findings. First, human performance is high - an average of a 1% word error rate for each of the three noise conditions. Second, human performance measured in terms of word error rates did not vary significantly with SNR. Machine performance on the same data (to be reported at this conference) is at best an order of magnitude worse and degrades very significantly at the lowest SNR of 10 dB.

1. INTRODUCTION

Human benchmarking, once an important aspect of any corpus collection [1-4], has fallen by the wayside in the last ten years in speech technology development. This is partially due to the fact that it is a well-accepted fact that human performance still significantly exceeds machine performance on a wide range of tasks. For example, one task on which human performance is well-documented is that of continuous digit recognition. Even though great strides have been made in digit recognition technology, human performance is still at least two orders of magnitude better than machines (the best machines perform at word error rates of about 1% on noisy telephone data).

The lack of solid human benchmarks is also partially due to the fact that such evaluations on extensive data sets are extremely expensive. Past evaluations have often focused on isolated word or syllable recognition. Rarely have such evaluations focused on fluently-spoken speech collected from data with an underlying large language model, and/or data characterized by extremely low signal-to-noise ratios (SNR). Hence, with the introduction of the Spoke 10 in the 1994 CSR evaluation, it is not surprising that the issue of inherent recognizability has resurfaced.

In this evaluation, our goal is to establish a benchmark for a human's ability to recognize and transcribe speech that will serve as a reasonable target for machine performance. Obviously, for such an upper bound, context plays an extremely important role in overall speech understanding performance. Since the evaluation corpus (described in detail below) consists of a group of sentences excised from the Wall Street Journal-based corpus [5], full context available at the time a reader first encounters such material (including the historical context in which it occurred) is not easily replicated for such an evaluation. However, we propose a methodology which we believe will support the reconstruction of the most significant amounts of contextual information, and yet minimize the amount of material required to conduct the evaluation.

2. SPOKE 10

Let us begin by briefly describing the Spoke 10 evaluation, referred to as a spoke, of the 1994 ARPA CSR evaluation [6]. The original speech material consists of nominally 11 utterances per speaker recorded from 10 different speakers (5 males and 5 females ranging in age from 18 to 59), and is a subset of the 1993 5K-word Wall Street Journal corpus (WSJ1). There are a total of 113 utterances in the actual corpus. The data for each speaker represents a paragraph-sized unit separated into a one utterance per file format. Hence, hypothetically, there is significant contextual information that can be brought to bear on the recognition process within a given speaker's data.

The data were originally recorded as read speech under high quality recording conditions using a Sennheiser HMD414 microphone. Spoke 10 consists of an evaluation on data degraded by adding noise at three SNR levels (unknown to the machine-based recognition systems): 22 dB, 16 dB, and 10 dB. The noise samples were collected in a Nissan Maxima automobile traveling at roughly 62 m.p.h. with all windows closed and the air-conditioning adjusted to a comfortable level. The noise was recorded using an omnidirectional microphone, a Shure SM-11CN, clipped to the driver's side sun visor.

A one minute sample of the noise for each noise level, preceding the noise segment actually added to the speech, is available for training purposes. A global speech level was computed based on a standard procedure defined by NIST [6]. Noise samples were subsequently digitally added to the original S0 data to achieve the prescribed SNRs. The total amount of data available for evaluation in Spoke 10 is 339 utterances: 113 utterances from 10 speakers times three noise levels. Note that the SNR per utterance varies depending on the vocal effort level of the speaker.

3. EXPERIMENTAL DESIGN

A successful evaluation must minimize the amount of time required to execute the evaluation, yet maximize the subject's ability to adapt to necessary nuances of the evaluation. In this case, we expected some listener adaptation to occur to normalize characteristics of the speakers, additive noise, and the inherent difficulty of the given material (a passage involving a topic unfamiliar to the listener). We primarily sought to calibrate performance as a function of the SNR, and to limit performance variations due to such second-order effects as those described above.

Key components of our testing methodology were:

- subjects were college-educated adults whose native language is English and who were familiar with various aspects of the Unix computing tools used in the evaluation;
- each subject listened to 113 utterances in a random order, and never heard an utterance from the same source prompt more than once;
- subjects were allowed to iterate over the entire dataset as much as is desired - no constraints were placed on the order in which utterances are transcribed or corrected, or the amount of time spent on the task;
- subjects were NOT be allowed to alter any of the reproduction conditions, EXCEPT the volume level of the each utterance — subjects were allowed individual control of the volume level of an utterance, but no other form of audio processing (i.e., a bass/treble control) will be allowed;
- subjects listened to the data via closed earcushion headphones using a high quality audio system (a 16-bit Sun Sparcstation 5 audio system);
- subjects transcribed the data using standard orthography aided by an interactive editor (emacs) and spelling checker (spell), and were allowed to modify/correct, etc. any transcription for any utterance at any time;

While this test was not designed to establish the absolute upper bound on human performance, we believe these results approximate the best possible human performance on such a task. Better performance could be obtained by adding more contextual information (for example, letting a single listener listen to all utterances from a given speaker in the order in which they appeared in the prompting text, and letting a single listener score as many utterances from the same speaker as possible).

Yet, we believe such a format for such limited amounts of data would not provide for a proper randomization of the data, thereby decreasing the accuracy of error rates by increasing their variance across listener and speaker (and nature of the material). In our proposed methodology, each listener evaluated approximately the same number of utterances for each noise condition, and had the opportunity to hear each speaker across all three noise conditions. We feel this methodology minimized the number of utterances each listener is required to evaluate, and minimized the number of listeners, without significantly jeopardizing the overall results.

3.1. Subject Selection

Of course, such evaluations depend heavily on the diligence of the subjects. Subjects with the following characteristics were selected based on *a priori* knowledge of their capabilities:

- native-born American citizens for whom English is their first and primary language;
- normal hearing (no known hearing-loss or other abnormalities);
- college-educated adults (a mixture of graduate students and faculty);
- (Unix) computer literate.

An overview of the demographics of the listener population is given in Table 1.

Listen	Sex	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

Table 1. An overview of the demographics of the listener population is shown, including the sex, age, number of years of post-high school education, number of sessions and amount of time required to complete the task, and the average uncorrected word error rate for the "no noise" condition (somewhat of a measure of the skill of the listener).

3.2. The Ambient Environment

All attempts were made to provide a quiet ambient environment for the evaluation. The room used during the evaluation was a terminal room with a fairly low background noise level (approx. 50 dBA). This room contained a number of X terminals and small Unix workstations, and was normally fairly quiet. During the period of the evaluation, the room was lightly populated and hence relatively quiet (but did contain several subjects performing the experiment as well as an occasional group of students working at the terminals). Listeners used high quality lightweight stereo headphones: Sony model number MDR-V50

valued at $25). These featured earcushions that encase the ear for a high degree of acoustic isolation from the ambient environment.

3.3. The D/A Audio Interface

The audio interface employed in this experiment was the standard Sun Sparcstation 5 16-bit D/A converter. The most recent Sparcstation 5 audio is by far the best audio interface Sun has delivered to date. Pilot experiments with this audio system (using an interface based on the public domain software package SOX and the standard Sun play command) indicated the audio system was adequate in quality. The only minor complication with the audio interface was that the speech data had to be digitally amplified prior to the experiment to insure there was sufficient headroom for the individual listeners' personal preference.

Listeners were allowed to adjust gain, but not any other audio parameter, in playing the data. Originally, we desired to use a single volume setting, adjusted to be a comfortable level for all utterances for all listeners. Our logic behind this was that we would like to preserve as much of the natural variation of amplitude in the data as possible (a similar argument can be made for restricting the use of bass and treble controls to preserve spectral balance).

Initial pilot experiments showed this policy was unacceptable to the pilot experiment listeners — there is sufficient variation in the level of the data, and a significant variation in the desired volume level for each listener. Constraining each listener to the same volume level proved to be unsatisfactory to the listener. Hence, we adopted our backup plan of providing a single volume control and allowing the listener to adjust this level for each utterance. (An alternative strategy that was suggested was to allow the listener to adjust the level at the beginning of the session. However, this also has some operational problems and could produce unreliable results if the listeners unknowingly set the volume level to the wrong value.)

3.4. User Interface Issues

Perhaps the most serious issue in conducting these evaluations involved the mechanics of the data entry task. With naive subjects, this can be a very real problem, since the typical user is not accustomed to transcribing what they hear. The chance of artifacts appearing in the data is great. Hence, we circumvented this problem by developing a plan to use knowledgeable subjects.

The next most serious issue was the level of detail of the transcriptions. Fortunately, since these are read speech items, and fairly clean data at that, we did not need to worry about transcription of dysfluencies, false starts, jargon, or complicated syntax. Listeners transcribed the data using normal orthography, indicating contractions as such (using apostrophes), possession, mixed-case, etc. The data was then converted off-line, under the guidance of the researchers, to the proper evaluation format [7].

Subjects were allowed the use a fairly powerful editor, the GNU free software tool emacs, and its associated standard spelling checker — spell. No grammar tools, or other word processing tools (such as a synonym tool), were allowed. Domain specific dictionaries, or word lists (such as a list of the 5000 word closed-set vocabulary) were NOT provided. A standard printed dictionary was also available for use. Subjects referred to files by number (001-113) and never saw the original filename or any other such identification information designating the noise level or speaker.

Subjects were NOT allowed to use graphical tools to display waveforms or spectrograms, and were not allowed to listen to portions of the utterances in isolation. The only audio presentation available was a function key that played the entire utterance. Subjects were allowed to play any file at any time, and to modify any corresponding transcription at any time (by traversing an emacs buffer containing all transcriptions).

Subjects participated in a training phase immediately prior to the formal evaluation. They were presented four utterances from the Spoke 10 development test set that included an original utterance, referred to as the "no noise" condition, and the corresponding version from each of the three noise conditions. Their transcriptions were checked to make sure there were no procedural errors, but the users were given no feedback about the accuracy of the transcription (unless there were procedural errors).

3.5. Stimuli Preparation

In addition to the three noise levels comprising Spoke 10, subjects were also presented a "no noise" condition. Some constraints on the randomization of the utterances is given below:

- each listener processes 113 utterances;
- each listener hears only one version of each original source utterance;
- each listener hears an equal number of examples from each noise condition;
- noise conditions and speakers are presented in a fairly random order throughout the session;
- a group of four listeners should cover the entire test set;
- a group of three listeners should be exposed to the same stimuli so that a single committee decision with a clear majority can be produced involving all listeners.

Given the limited amount of data and resources available for this experiment, we designed stimuli based on a set of 12 listeners: 3 groups of four listeners. Each utterance in the 113 x 4 condition corpus was heard by exactly three listeners. Further, a group of four listeners combined to form a single evaluation of the entire dataset for Spoke 10. Hence our results can be collated to provide three separate benchmarks for Spoke 10 that can be compared directly to machine performance.

4. EVALUATION

Despite the fact that Spoke 10 is a closed-vocabulary evaluation, we choose to conduct our experiment as an open-vocabulary test. Spelling corrections were then applied as a postprocessing step to correct out-of-vocabulary words. In Table 2, we summarize the overall performance of these listeners for these two conditions. The results are averaged across all noise conditions. The second row, labelled "Committee," corresponds to a machine-generated committee decision in which three listeners' data for each utterance was pooled on a word-by-word basis to form a common transcription. We expect this condition to remove most of the

Evaluation Group	Vocabulary	
	Open	Closed
Average	2.1	1.0
Committee	1.2	0.5

Table 2. An overview of human performance on the Spoke 10 corpus is given. The columns correspond to the open-set and closed-set vocabulary conditions, while the rows correspond to the average performance across all listeners, and a committee decision based on groups of three listeners.

errors due to inattention by the listeners (unintentional errors), and to resolve some ambiguities due to particular listeners' unfamiliarities with the task.

Listeners' familiarity with the specific domain and its specialized terms (primarily proper nouns) played a significant role in the overall performance. The error rate dropped by a factor of 2 when spelling corrections (based mainly on misspellings of proper nouns) were applied (for example, "Fannie Mae" was substituted for "Fanny May" and "Shearson Lehman" replaced "Sheerson Leeman"). Yet, even without such corrections, performance is still almost an order of magnitude better than the best machine performance (to be reported at this conference).

4.1. The Open Vocabulary Evaluation

A more detailed analysis of the open vocabulary evaluation is given in Table 2. Each table entry below corresponds to a listener (or group of listeners) and a noise condition. Recall that each listener transcribed 113 utterances: approximately 28 from each noise condition (including the no noise case). Hence, the score reported for "ALL" consists of approximately the same number of utterances from each condition.

Also recall that 4 listeners combine to form one group (one utterance heard once for all noise conditions). This is the result that most closely matches the formal definition of Spoke 10. Hence, groups 1, 2, and 3 evaluated the same data, and these results can be compared directly. Finally, note that a difference of 0.1% is approximately the equivalent of a difference of one word error for most conditions, so that, for most of the results below, the differences are not statistically significant.

An interesting statistic to look at is the amount of agreement among listeners. 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 disagreed, a transcription was arbitrarily chosen (by rule the transcription corresponding to the first speaker in the group).

There were 179 word occurrences that accounted for the errors in the listeners' data. Of these 179 words, 105 (59%) fell outside of the 5,000 word vocabulary defined for the Spoke 10. The errors could be split into four groups: misspelled proper nouns (63%), misuses of possessives (7%), contractions (3%) and other (for example, "chairmen" for "chairman" and "jester" for "gesture").

Listener	Noise Condition				
	None	22 dB	16 dB	10 dB	All
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

Table 3. An overview of human performance for the open-vocabulary condition. We observe performance is extremely high, and does not vary significantly with SNR.

4.2. An Analysis With Spelling Corrections

The data was postprocessed with a set of spelling corrections derived by replacing words outside the language model with their homophones inside the language model. This formed what we refer to as a "closed-vocabulary" test. It is not a true closed-vocabulary test because the subjects were not asked to make the judgements with a priori knowledge of the language model (which would have influenced their results). Nevertheless. the spelling corrected data forms an important baseline for human performance. The results are given in Table 4.

A group of three listeners, which formed the committee decision, disagreed on only approximately 25% of the utterances. ALL except two of these could be resolved into a single transcription by a simple majority decision. In one of these cases, the error was possibly the result of inattention. In the other case, the error involved a substitution between articles "the" and "a."

It is also interesting to note that, after spelling correction, only 13% of the sentences were in error (before spelling corrections, 25% of the sentences were in error). Sentence errors were not a function of the length of the sentence, demonstrating the power of context in recognition. For the committee decision, only 6.6% of the sentences were in error. The most common error modalities were equally distributed amongst all standard categories.

A detailed analysis of the errors exposed the fact that at least half of these errors were probably due to inattention. Given that we were not using trained transcribers, and that the subjects were not being given incentives to perform well, this is not surprising. Since such an error can often modify the error rate by at least 0.1%, these errors have a profound impact on our ability to perform detailed analyses.

Listener	Noise Condition				
	None	22 dB	16 dB	10 dB	All
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

Table 4. The results for the spelling-corrected data show that the error rate approximately halves, both for the individual listeners and the committee decision.

4.3. Removal of Duplicates

Spoke 10 was not the ideal corpus for this type of experiment. One of its deficiencies is the duplication of prompting material across speakers (several speakers said the same thing). Hence, some subjects heard the same prompting material more than once. Theoretically, a subject's performance on the utterance should improve the more times it is presented (in this case, by different speakers). To offset this, we performed an analysis in which all utterances which were prompted more than once were removed. We refer to this as the "no duplicates" condition. The results are presented in Table 5.

The "no duplicates" results are not significantly different from the results in Table 4. Though we had hoped to see an increased sensitivity to SNR, this does not appear. We believe this is because the SNR was simply not low enough. Had the SNR been lower, perhaps as low as 5 dB, a greater sensitivity may have resulted. The data, even at 10 dB SNR, was highly recognizable — the listeners did not appear to have much difficulty transcribing it.

Since the noise-degraded stimuli were generated using a single global speech level estimate, and hence the SNR per speaker could fluctuate, we also analyzed performance as a function of speaker for the "no-duplicates" condition. There was no strong correlation with performance as a function of speaker and noise level. A summary of the error rates per speaker, averaged across all listeners, is shown in Table 5.

4.4. Error Analysis

Our final analysis consists of taking the spelling-corrected "no duplicates" data and sorting the errors on basis of the type of error

Listener	Noise Condition				
	None	22 dB	16 dB	10 dB	All
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

Table 5. The performance after removing duplicate prompts is not significantly different from the results with duplicate prompts.

Speaker	Noise Condition				
	None	22 dB	16 dB	10 dB	All
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

Table 6. The performance as a function of speaker, averaged across all listeners, is shown.

modality. This analysis is given in Table 5. There are three classes of errors: inattention, an anomalous problem with an utterance containing "the the," and valid auditory-based transcription errors. The latter category was further subdivided into the number of phone errors that resulted from the corresponding word errors. From this data, we see that a large portion of the sentence errors involve a single morpheme-sized unit (often a function word).

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
5 phones	1 (1%)

Table 7. The performance as a function of the nature of the error. There are three major categories: inattention, an anomalous case where the prompt contained “the the” and was consistently transcribed as “the,” and valid acoustically-motivated transcription errors. The latter are sorted by the number of phone errors that occurred as a result of each error. From this data, we see that a large percentage of sentence errors were caused by one and two phone confusions (often an error in a suffix, prefix, or article).

Finally, in Appendix A, a list of the sentence errors for the transcriptions derived from a machine-based committee decision of the spelling corrected data are given. For this evaluation there were a total of 30 sentences with errors. Fourteen different prompts were responsible for these 30 errors — a ratio of 2 errors per prompt. Observe how consistent the error modalities were across the various noise conditions. Appendix A further supports our conclusion that humans transcribed the noisy data without serious difficulty, and shows the small effect the additive noise had on the semantically important portions of the sentence.

5. SUMMARY

This experiment resulted in two important findings. First, human performance is high — an average of a 1% word error rate for each of the three noise conditions. Machine performance on the same data (to be reported at this conference) is an order of magnitude worse.

Second, human performance did not vary significantly with SNR. This speaks well of humans’ ability to separate a speaker from noise when noise has been digitally added in an uncorrelated fashion. Clearly, the differing spectral and temporal structures of the two signals are keys to allowing humans to separate the signals at some level in their speech understanding systems. Context also obviously plays a role, though in this study we have shown very little context is needed.

The fact that performance did not vary significantly with SNR suggests two things. First, performance does not gracefully degrade with SNR. It seems there is a sharp threshold somewhere in the range of 0 dB to 10 dB (depending on how one measures SNR). Second, human performance is most likely at least 10 dB ahead of machine performance in SNR. Machine performance tends to degrade rapidly in the 10 dB range. Human performance, especially on uncorrelated noise, appears unaffected by SNRs as low as 10 dB.

6. ACKNOWLEDGEMENTS

First, and foremost, we thank the twelve listeners who donated their time to this study. This was a tedious task that required almost three hours to complete. Without each subject’s hard work, and excellent workmanship, we would not have completed this study in time for the CSR’94 evaluations.

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. “The gang at NIST” — particularly John Garofolo, Jon Fiscus, and Bill Fisher provided a large amount of tactical support, and spent a great deal of time consulting with us on this project. Finally, we thank Ron Cole at the Center for Spoken Language Understanding for his contributions to the experimental design (and successfully predicting its outcome) and Raja Rajasekaran of Texas Instruments for providing valuable feedback on the experimental design.

REFERENCES

1. R.G. Leonard, “A Database For Speaker-Independent Digit Recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 42.11.1 - 42.11.4, San Diego, California, USA, April 1984.
2. Y.K. Muthusamy and R. A. Cole, “Automatic Segmentation and Identification of Ten Languages Using Telephone Speech,” in *Proceedings of the 1992 International Conference on Spoken Language Systems*, pp. 1007-1110, Banff, Alberta, Canada, October 1992.
3. R. Cole, B.T. Oshika, M. Noel, T. Lander, and M. Fanty, “Labeler Agreement in Phonetic Labeling of Continuous Speech,” in *Proceedings of the 1992 International Conference on Spoken Language Systems*, pp. 2131-2134, Yokohama, Japan, September 1994.
4. G.N. Tajchman and M. A. Bush, “Effects of Context and Redundancy in the Perception of Naturally Produced English Vowels,” in *Proceedings of the 1992 International Conference on Spoken Language Systems*, pp. 839-842, Banff, Alberta, Canada, October 1992.
5. D. B. Paul, J. M. Baker, “The Design for the Wall Street Journal-based CSR Corpus,” in *Proceedings of the 1992 International Conference on Spoken Language Systems*, pp. 899-902, Banff, Alberta, Canada, October 1992.
6. D. Pallett, “Draft Proposal for the 1994 CSR Evaluation,” The National Institute of Standards and Technology, Room A216 Building 225 (Technology), Gaithersburg, MD 20899, October 5, 1994 (available by ftp from jaguar.ncsl.nist.gov). (Also, to appear in the *Proceedings of the 1995 ARPA Human Language Technology Workshop*, Austin, Texas, USA, January 1995.)
7. D. Pallett, “System Output Preparation and Scoring Protocols,” The National Institute of Standards and Technology, Room A216 Building 225 (Technology), Gaithersburg, MD 20899, October 4, 1994.

APPENDIX A: A List of All Errors For Committee Transcriptions

Note that only one specific recognition error was made for any source utterance, and that this error was made, on the average, in over half of the various noisy versions of each utterance. This suggests that the errors might be attributable in part to imperfect articulation on the part of the speaker or to judgements made in creating the reference transcription.

		Noise Level			
ID	Transcriptions: (R) Denotes Reference; (H) Denotes Human Hypothesis	Clean	22 dB	16 dB	10 dB
1. 4T0C0304	(R) the **INDEX HAS** averaged fifty four %percent... (H) the **INDEXES ***** averaged fifty four %percent...	x	x	x	x
2. 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
3. 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
4. 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
5. 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
6. 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	
7. 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
8. 4TCC0301	(R) ...shareholder approval for the **PLAN** at its annual meeting... (H) ...shareholder approval for the **PLANT** at its annual meeting...	x			
9. 4TGC0306	(R) we **CAN** compete (H) we **CAN'T** compete			x	x
10. 4TCC030A	(R) ...all the economic indicators are solid **"QUOTE** and he attributed ... (H) ...all the economic indicators are solid **QUOTA** and he attributed ...	x			
11. 4TCC0301	(R) ...and nuclear technology concern **SAID** it would seek shareholder... (H) ...and nuclear technology concern **SAYS** it would seek shareholder...			x	
12. 4T2C0304	(R) the fully diluted figure reflects **A** forty .point three million... (H) the fully diluted figure reflects **THE** forty .point three million...				x
13. 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
14. 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