BENCHMARKING HUMAN PERFORMANCE FOR CONTINUOUS SPEECH RECOGNITION

Neeraj Deshmukh, Richard Jennings Duncan, Aravind Ganapathiraju, Joseph Picone

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

ABSTRACT

It is a well-established fact that human performance exceeds that of computers by orders of magnitude on a wide range of speech recognition tasks. However, there is widespread belief that the gap between human and machine performance has narrowed considerably on restricted problems. Yet, there are few extensive comparisons of performance on tasks involving large vocabulary continuous speech recognition (LVCSR) and low signal-to-noise ratios (SNRs). Human evaluations on LVCSR tasks highlight a number of interesting issues. For example, familiarity with the domain plays a crucial role in human performance.

We conducted several experiments that extensively characterize human performance on LVCSR tasks over two standard evaluation corpora — ARPA's CSR'94 Spoke 10 and CSR'95 Hub 3. We demonstrate that human performance is at least an order of magnitude better than the best machine performance, and that human performance is fairly robust to a number of factors that typically degrade machine performance: SNR, speaking rate and style, microphone and ambient noise. In fact, human performance remained remarkably consistent across evaluation paradigms, and to some extent was artificially limited by a listener's attention span.

1. INTRODUCTION

Automatic speech recognition technology has made significant advances in the past decade — current state of the art systems are capable of performing continuous speech recognition (CSR) over complex domains and large vocabularies involving tens of thousands of words. A significant part of this progress can be attributed to the development of large speech corpora used to train such systems. It is well-accepted that human recognition performance still significantly exceeds the best machine performance, but recent attempts to calibrate this difference have not focused on state-of-the-art performance tasks [1].

A possible reason for the lack of good human benchmarks is that such evaluations on large vocabulary tasks are extremely expensive. Moreover, when conducting such tests it is difficult to replicate the full contextual information in the test data source when a participant reader first encounters the material. This is a significant problem as context is vitally important in overall speech understanding.

A better understanding of how humans perceive speech under different conditions will also facilitate recognition using machines. Therefore, apart from setting a goal for machine error rates, another motivation to conduct evaluations for humans is to study the effects of various factors on human recognition performance; such as the noise content in speech, the artifacts introduced by the placement and type of microphone used for recording the data, characteristics of the speaker (accent, speaking rate etc.) and the properties of the spoken material (e.g. complexity, vocabulary set).

With these objectives in mind, we conducted a number of experiments to establish a benchmark for human recognition performance on the data used in the 1994 and 1995 ARPA CSR evaluations [2, 3]. By systematically distributing the test data among the listeners, and by using both native and non-native speakers of English as subjects, we have managed to achieve a very close approximation to ultimate human performance with a paradigm that minimized the resource requirements.

2. EVALUATION CORPORA

The 1994 Spoke 10 corpus (94-Spoke 10) [4] was created to study the effect of noise on the performance of speech recognition systems. The 1995 Hub 3 corpus (95-Hub 3) [5] was designed to calibrate the effect of microphone type and placement on recognition performance. 95-Hub 3 also indirectly assesses recognition of proper-nouns and out-of-vocabulary (OOV) words — issues which pose very interesting challenges for human subjects (is the listener familiar with particular jargon?).

Both corpora consist of speech data recorded from a variety of speakers (equal number of males and females, all native speakers of American English). The data for each speaker represents a contiguous paragraph-sized unit, thus providing systems with the opportunity to utilize meaningful contextual information.

94-Spoke 10 Corpus: This corpus consists of 113 sentences drawn from a subset of the 1993 5K-word Wall Street Journal (WSJ1) corpus [4]. These transcripts were recorded from 10 speakers (typically 11 sentences per speaker) using a Sennheiser HMD-414 microphone. The data was degraded by adding noise at three SNR levels (22 dB, 16 dB and 10 dB) to create a total of 452 utterances.

95-Hub 3 Corpus: This corpus contains 20 paragraphs, each 15 sentences in length, drawn from articles appearing in a broadcast news program [5]. Each utterance was recorded using two microphones, making a total of 600 utterances. One of the two microphones — a close-talking Sennheiser HMD-410 (*mic s*) — was fixed for all speakers. The second microphone was one of 3 alternates — a boom-mounted Shure SM58 (*mic b*), an Audio Technica AT851a Micro Cardioid (*mic f*), and a Radio Shack 33-1060 Omni Electret in a slip-on desk stand (*mic g*).

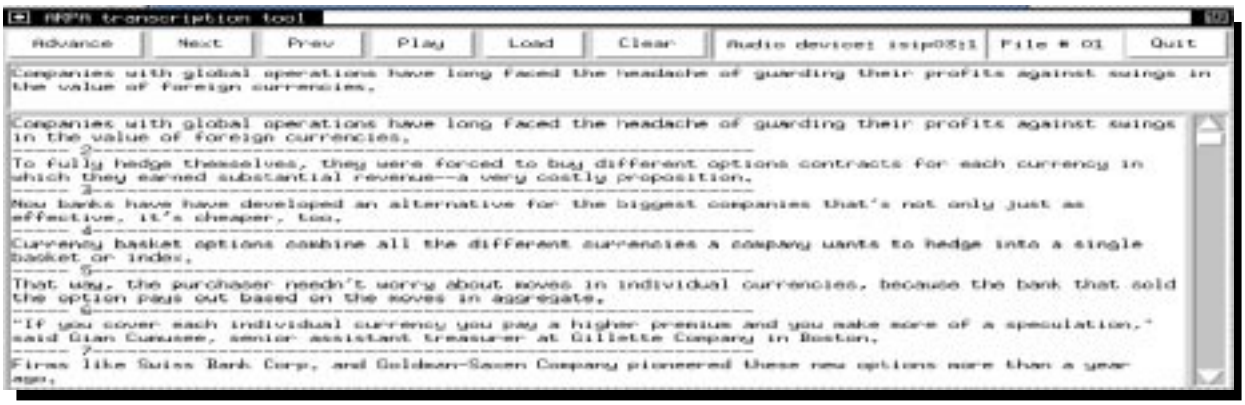

Figure 1: Transcription tool interface for the human benchmark evaluations

3. EXPERIMENTAL DESIGN

A successful evaluation must minimize the resources required to execute the speech recognition experiments, yet maximize the listeners' ability to adapt to the necessary nuances of the task.

With this approach in mind we devised an experimental setup in which every listener

- was a college-educated adult, computer literate (UNIX-based) and a native-born American with English as the primary language (except two listeners on 95-Hub 3 who were non-native speakers of English);
- listened to a subset of the data (for Hub 3 containing 120 sentences, 60 on *mic s* and 60 on a mixture of other microphones; for Spoke 10 — 113 sentences distributed evenly over all four noise conditions) in random order;
- was allowed to iterate over the entire data subset as much as desired during transcription;
- listened to the data using a 16-bit audio system and closed ear-cushioned headphones;
- was not allowed to alter any of the playback conditions except volume (playing short segments of an utterance, or looking at waveforms/spectrograms of the data, was not permitted);
- used normal orthography for transcription aided by a Tcl-based transcription GUI [Figure 1].

Thus, each listener evaluated approximately the same number of utterances for each condition, and was exposed to a different condition for each speaker (for a given listener, no speaker appeared more than once). This methodology minimized the number of utterances each listener was required to evaluate, and minimized the number of listeners required to perform a comprehensive evaluation, without significantly jeopardizing the overall results. Twelve listeners (6 males/6 females) for 94-Spoke 10 evaluations were used, while 15 listeners (8 males/7 females) were used in 95-Hub 3. Two listeners were common to both the tests, providing a direct comparison of the level of difficulty of each task.

3.1. Benchmarking Methodology

Given the limited amount of data and resources available for this experiment, we split the subjects into three groups for each of the two benchmarks. Each utterance (under each condition) in the corpora was transcribed by exactly three listeners, thus providing a *committee transcription*. Moreover, every group of listeners combined to form a single evaluation of the entire data set for each corpus. Thus our results provide five separate benchmarks (three groups, overall and committee) that can be compared directly with the machine performance.

The committee transcriptions, created by pooling the data for each utterance from all three listeners on a word-by-word basis, are important as they remove most of the errors due to inattention by the listeners (unintentional errors), and resolve some ambiguities due to particular listeners' unfamiliarity with the task.

The transcriptions were conducted as an open-vocabulary test, with the listeners given access to the lexicon spanning the language model for each corpus. A major factor affecting listener performance was the attention span — most listeners committed numerous errors (e.g. misspellings) due to inattention. Therefore as a postprocessing step, the transcriptions were converted into an augmented-vocabulary set by applying spelling corrections. Well-known proper nouns that had incorrect spellings were replaced with their counterparts in the language model (typically names we felt a good speller would know, such as "Tokyo"). The error rate dropped considerably with such corrections [Figure 2].

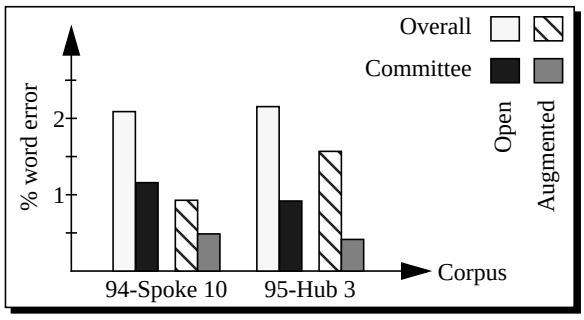

Figure 2: Overview of human recognition performance

Listener	% Word Error on Noise Condition				
	None	22 dB	16 dB	10 dB	All
Group 1	1.8	2.0	2.0	1.6	1.9
Group 2	1.8	2.2	1.9	2.0	2.0
Group 3	2.5	2.2	2.5	2.7	2.5
Overall	2.0	2.1	2.1	2.1	2.1
Committee	1.0	1.4	1.2	1.2	1.2

(a) 94-Spoke 10 Corpus

Listener	% Word Error on Microphone Condition				
	Mic s	Mic b	Mic f	Mic g	All
Group 1	2.1	2.3	2.0	3.0	2.3
Group 2	1.4	1.3	0.9	2.2	1.4
Group 3	2.5	2.1	2.8	4.7	2.8
Overall	2.0	1.9	1.9	3.3	2.2
Committee	0.8	1.0	0.2	1.5	0.8

(b) 95-Hub 3 Corpus

Table 1: Overview of the open vocabulary evaluation for human recognition performance

Listener	% Word Error on Noise Condition				
	None	22 dB	16 dB	10 dB	All
Group 1	0.9	0.7	0.7	0.9	0.8
Group 2	1.0	0.9	0.7	1.0	0.9
Group 3	1.2	0.7	1.0	1.5	1.1
Overall	1.0	0.8	0.8	1.1	0.9
Committee	0.6	0.3	0.4	0.7	0.5

(a) 94-Spoke 10 Corpus

Listener	% Word Error on Microphone Condition				
	Mic s	Mic b	Mic f	Mic g	All
Group 1	1.7	1.7	1.4	1.9	1.7
Group 2	1.0	1.0	0.8	1.5	1.0
Group 3	1.9	1.5	2.6	3.3	2.1
Overall	1.5	1.4	1.6	2.2	1.6
Committee	0.3	0.6	0.1	0.8	0.4

(b) 95-Hub 3 Corpus

Table 2: Overview of the augmented vocabulary evaluation for human recognition performance

4. EVALUATION RESULTS

Listeners typically required 2 to 5 hours to complete each of the two evaluation tasks. Most listeners chose to transcribe the data in a single session, the remainder needed two. Listeners' familiarity with the specific domain and its specialized terms played a significant role in their overall performance. Each group of listeners heard each utterance once for all noise/microphone conditions. Thus groups 1, 2 and 3 evaluated the same data and their results are directly comparable. It was observed that an error of 0.05% is the equivalent of a single word error for most conditions; thus for most of the results the differences are statistically insignificant.

4.1. Open Vocabulary Evaluations

Table 1 provides a summary of the open vocabulary test results. The error rate was found to be more or less consistent across all groups and all noise conditions — an indication that humans do not find these variations particularly challenging. This constant level of performance is significant considering the fact that machine performance degrades rapidly in such environments.

Examination of the agreement among listeners yields some interesting statistics. A group of three listeners who transcribed the same data 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% generally pertained to proper nouns. In rare cases where all three listeners had a disagreement the transcription closest to the truth was selected. The overall sentence error rate was 25% on 94-Spoke 10 and 28% on 95-Hub 3, while the committee results were 17% and 13% respectively.

4.2. Augmented Vocabulary Evaluations

The results for the augmented vocabulary evaluation are given in Table 2. There was a significant decrease in the word error rate, reflecting the importance of the problem of proper-noun recognition. Overall sentence errors dropped to 13% and 22% respectively (7% each for committee transcriptions) on the two corpora, and were independent of the sentence length, thus demonstrating the power of context in recognition.

The committee disagreed on only about 25% of the utterances, most of which were resolved by a simple majority rule. A large percentage of word errors appear to be motivated by inattention rather than by unfamiliarity with the subject matter. The word error rate dropped 80% from the group case to the committee decision.

Given that we did not use trained transcribers, and that the subjects were not given any incentive to perform well, the drastic reduction in error rate from the groups to the committee is not surprising. However, such anomalous errors have a profound impact on our ability to perform detailed analyses. The difference in the committee decision and the individual group transcriptions would shrink if we were to use highly-trained transcribers.

5. ERROR ANALYSIS

No significant correlation was found between human recognition performance and SNR levels or microphone conditions. This speaks well of humans' ability to separate a speaker from noise when noise is additive and uncorrelated. Evidently, the differing spectral and temporal characteristics of speech and noise are keys that allow humans to separate the two signals at some level in their speech understanding systems. Performance was also found to be largely insensitive to microphone characteristics (only the Omni Electret, *mic g*, posed a few problems for the transcription).

Other factors affecting recognition performance were context — the listeners' familiarity with the material and speaker characteristics. Performance degraded somewhat with faster speaking rates. On 95-Hub 3 there was a particular speaker who was much harder to transcribe due to poor articulation.

Transcription	Utterance with error
Reference	...the **INDEX HAS** averaged fifty four percent...
Hypothesis	...the **INDEXES ***** averaged fifty four percent...
Reference	**YOU** could learn a lot from a so-called slacker...
Hypothesis	**WE** could learn a lot from a so-called slacker...

Transcription	Utterance with error
Reference	...chips that store two hundred **AND** fifty six...
Hypothesis	...chips that store two hundred ******* fifty six...
Reference	...cut back holdings **OF** public money managers...
Hypothesis	...cut back holdings **IN** public money managers...

Table 3: Characteristic examples of human transcription errors for the committee evaluations on the augmented vocabulary test

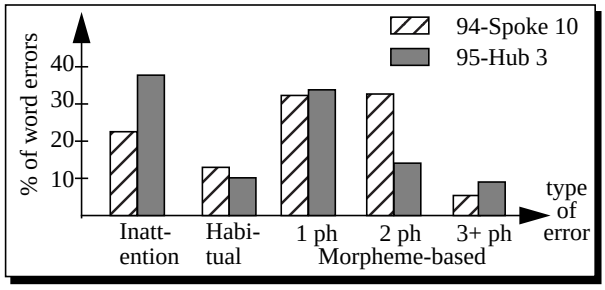

Figure 3: Common error modalities on the listening tests

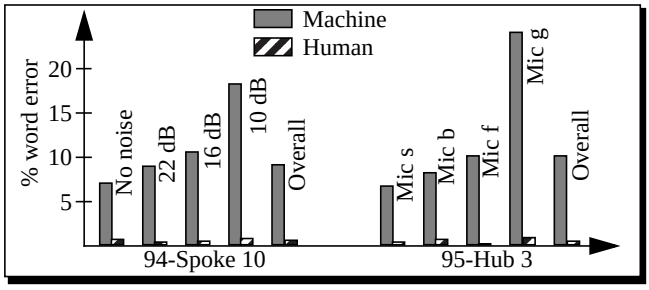

Figure 4: Comparison of human recognition performance with the best machine performance (results for the Cambridge HTK-based system [6, 7]) as a function of SNR and microphone conditions

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 subdivided the latter category into the number of phone errors that constituted each word error [Figure 3]. 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. Some typical transcription errors are illustrated in Table 3.

Figure 4 presents a direct comparison of the human recognition performance with the best machine word error rate for the augmented vocabulary task. Clearly, humans outperform the best machines by at least an order of magnitude. The error rate without such spelling corrections, in spite of being lower by a factor of two than that on the augmented test set, was still an order of magnitude better than the best machine performance.

6. SUMMARY

This experiment resulted in two important findings. First, for large vocabulary continuous speech recognition human performance is at least an order of magnitude higher than the corresponding machine performance. Second, human performance is consistent over a range of SNR values and microphone conditions. Humans seem to perform consistently over time and over different domains. Performance of the listeners common to both corpora is comparable, despite significant differences in the perceived difficulty of specific tasks. The 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 for future benchmarks.

A large percentage of the errors 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.

7. REFERENCES

1. R.A. 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.

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

3. N. Deshmukh, A. Ganapathiraju, R.J. Duncan and J. Picone, "Human Speech Recognition Performance on the 1995 CSR Hub 3 Corpus," in *Proceedings of the SLST Workshop*, pp. 53-59, Harriman NY, February 1996.

4. 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.

5. D.B. Pallett, "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.

6. D.B. Pallett, et. al., "1994 Benchmark Tests for the ARPA Spoken Language Program," in *Proceedings of the SLST Workshop*, pp. 5-36, Austin TX, January 1995.

7. D.B. Pallett, et. al., "1995 Hub-3 NIST Multiple Microphone Corpus Benchmark Tests," to appear in *Proceedings of the ARPA Speech Recognition Workshop*, Harriman NY, February 1996.