# 2000 NIST EVALUATION OF CONVERSATIONAL SPEECH RECOGNITION OVER THE TELEPHONE: ENGLISH AND MANDARIN PERFORMANCE RESULTS

*Jonathan G. Fiscus, William M. Fisher, Alvin F. Martin, Mark A. Przybocki, David S. Pallett*

National Institute of Standards and Technology (NIST)
Information Technology Laboratory (ITL)
Room A216 Building 225 (Technology)
Gaithersburg, MD 20899
E-mail: alvin.martin@nist.gov

## ABSTRACT

This paper documents the use of conversational telephone speech test materials in the NIST coordinated evaluation conducted early in 2000. The primary evaluation was of General American English speech, but a subsidiary evaluation of Mandarin speech was also offered.

The primary test data consisted of twenty conversations collected for the original Switchboard Corpus but not released with the published corpus and twenty conversations from the CallHome English Corpus.

The lowest English word error rates this year were 19.3% for the Switchboard-type data and 31.4% for the CallHome data. These are considerably lower error rates than those achieved in previous evaluations, the most recent of which was in 1998. These error rate reductions were due in part to improved recognition systems, but also in large part to these test sets being easier than those used in previous evaluations. We discuss in the Appendices some reasons for these test sets being easier than previous test sets.

## 1. DATA

As in previous years, the English evaluation had two main parts, each using conversations from different corpora. The evaluation this year also included a third "noncompetitive" part as well. This third part used data from the Switchboard-1 Corpus for which reference transcription at the phone level had been created, and participating systems were required to submit both word level and phone level hypothesized transcriptions. This part of the evaluation is described elsewhere in these Proceedings [1].

The test data for the two main parts of the English evaluation were as follows:

- Twenty conversations collected along with the original Switchboard-1 Corpus but not included in the published corpus. These were among conversations set aside with the intention that they could later be used as heretofore-unseen data. Of the forty speakers in these conversations, thirty-six appear in conversations of the published Switchboard Corpus. It was recognized beforehand that participating sites would have thus used speech of these speakers in their acoustic training data, but it was felt that this would have a limited effect in terms of enhancing performance. (See the discussion of this issue in section 5 and in Appendix 1.)
- Twenty conversations from the CallHome English Corpus. These were among the conversations collected for this corpus but not previously released as they were intended for use in evaluations.

For each conversation, participating systems were given a sequence of "turns" to process, with a total duration of about five minutes. Each turn contained speech of only one of the conversants, represented in single channel mu-law form. The segmentation into turns this year was done using software provided by the Institute for Signal and Information Processing at Mississippi State University. This produced somewhat different turn boundaries from those used in previous evaluations, which were provided by the LDC (Linguistic Data Consortium) using previously developed BBN software.

The training data for this evaluation consisted of the following:

- The entire SwitchBoard-1 Corpus as previously released
- The first 100 conversations of the CallHome English training corpus
- The 60 CallHome English conversations used as test data in the 1996, 1997, and 1998 evaluations
- The entire Switchboard-2 Phase-1 Corpus (most of which, however, has not been transcribed)

In addition, sites could use for training any English data from other corpora publicly available at the time results were reported.

The development data for this evaluation consisted of the CallHome English and Switchboard-2 Phase-2 test data sets from the September 1998 evaluation. Each of these contained twenty conversations.

## 2. PARTICIPANTS

There were five participating sites in the English evaluation. These were:

- AT&T Labs-Research
- BBN Technologies (a part of GTE)

- Cambridge University Engineering Department (HTK group)
- Mississippi State University - Institute for Signal and Information Processing (ISIP)
- SRI International

Sites could submit results for several different systems, but one of them had to be specified as the site's primary system. Descriptions of the various systems may be found in other papers in these proceedings. It may be noted that the Mississippi State system was essentially the ISIP public domain automatic speech recognition system. In addition, as in past evaluations, a NIST-Rover [2] voting system was created combining the primary systems from the above five sites.

Each site estimated the processing time required by its primary system on a single processor of the type used in the evaluations. These estimates ranged from about 250 to about 820 times real-time.

There was one participant in the Mandarin evaluation, namely BBN Technologies.

## 3. RULES AND PROCEDURES

The evaluation rules and procedures were quite similar to those in previous evaluations (except for the "noncompetitive" part of the evaluation mentioned above). Full details may be found in [3].

Scoring, using NIST's *sclite* software package, was performed by aligning the system output for each turn with the turn reference transcription and then computing the overall word error rate (WER), which is defined as the number of words in error divided by the total number of reference transcription words. The alignments find three types of word errors:

- Substitution errors occur when aligned reference and system output words differ
- Deletion errors occur when reference words have no corresponding system output words
- Insertion errors occur when system output words have no corresponding reference words

The reference transcriptions consisted of a single sequence of words for each turn representing the transcriber's best judgement of what the speaker said. They were intended to be as accurate as possible, but there were necessarily some ambiguous cases and also some outright errors. The use of multiple human transcribers or a formal adjudication procedure was not regarded as cost effective in light of the current high error rates of automatic recognizers on this type of data.

The system output transcriptions could include any of several specified hesitation sounds such as "uh" and "um". For scoring purposes, all hesitation sounds were regarded as equivalent, with a common symbol ("%hesitation") used in the reference transcriptions. Moreover, omissions of hesitation sounds were not counted as deletion errors.

Systems were also asked to provide a confidence score along with each output hypothesized word. These scores were to represent estimates of the probability (in the range [0,1]) of the word being correct. While this could be merely a constant probability, certain applications and operating conditions could derive significant benefit from a more informative estimate sensitive to the input signal. A normalized cross entropy measure was computed as a measure of the usefulness of the confidence scores. For more details, see [4].

## 4. RESULTS

Figure 1 shows the word error rates for the primary systems from each site for both the Switchboard and CallHome portions of the test data. The lowest site word error rates, which were 19.3% for the Switchboard type conversations and 41.4% for the CallHome conversations, were achieved by the Cambridge group. The NIST-Rover system, combining the submissions of the other systems, achieved a slight improvement over this best score on the CallHome data and no improvement over this on the Switchboard type data.

As in previous evaluations, the error rates for the Switchboard type data were considerably lower than the CallHome data error rates. This is presumably due primarily to the different nature of the conversations involved. The Switchboard-1 speakers did not know each other before their conversations and generally adhered to the topics they were assigned. The CallHome speakers were generally family and friends of one another and spoke about whatever they chose. The greater topicality and formality of the Switchboard conversations aided recognition performance.

Figure 2 shows a scatterplot for the Cambridge system word error rates for the two sides of each conversation from each of this year's test corpora. Note that for this year's Switchboard conversations there is little correlation of the error rates for the A and B channels of each conversation ($R^2 = 0.02$). However, for the CallHome conversations there is considerable correlation ($R^2 = 0.37$) of the error rates for the two channels.

For the CallHome Corpus the speakers are family and friends of one another, and they are likely to share speaking styles and characteristics, such as accent, that can have a major impact on recognition performance. The Switchboard speakers, in contrast, did not know one another. It may be noted that listening to the one CallHome conversation whose two sides had the highest word error rates over all conversation sides revealed heavily accented speech and a very rambling conversation relating to people and events already familiar to the conversants.

It may also be noted in this context that analysis (by William Fisher) finds that in Switchboard the A channel initiator is somewhat less likely to use voiced hesitations than the B channel responder ($p < .007$) and a little less likely to have speech transcribed with optional elements ($p < .05$). No such trends in CallHome are noted. These observations are consistent with there being differences in

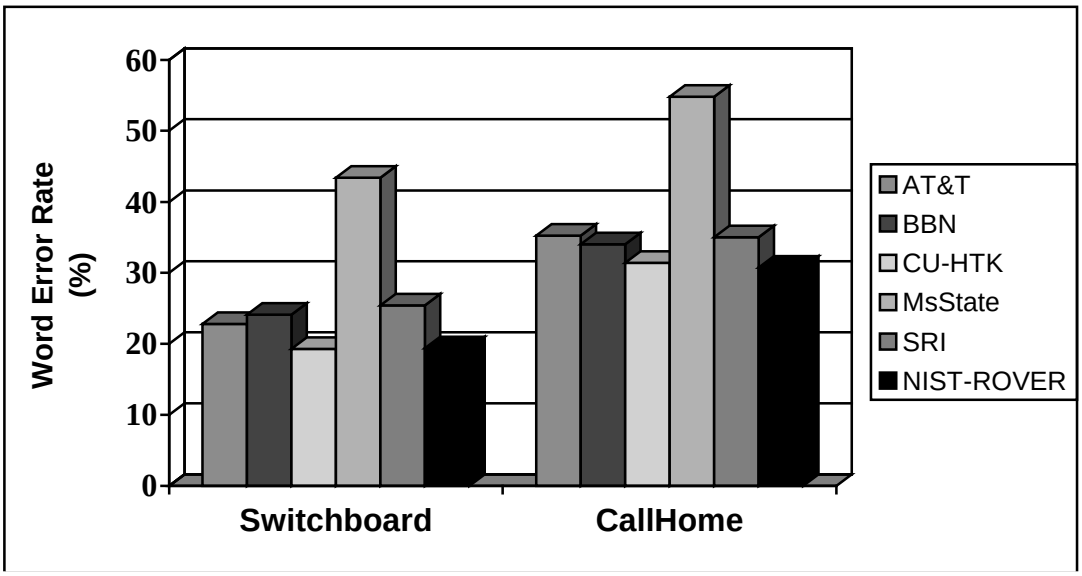


**Figure 1**: Word error rates of primary systems for each site

the word error rates for the two sides of the Switchboard conversations to a greater degree than in the CallHome conversations.

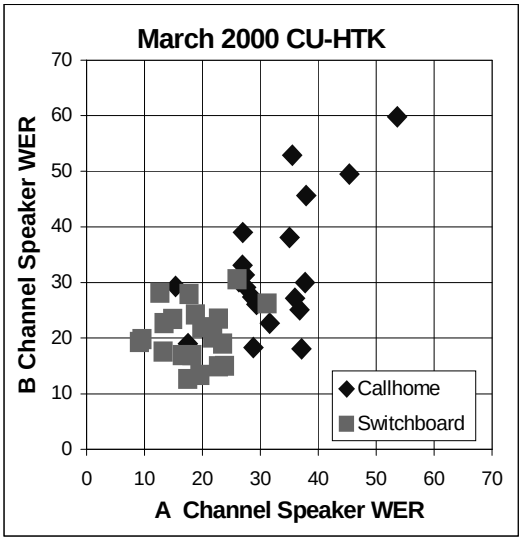


**Figure 2**: Scatterplot of the word error rates of the two sides of each conversation

NIST has long made it standard practice to examine the statistical significance of the performance results differences it measures in speech recognition evaluations. Several different tests are used, each examining, for each pair of systems considered, whether the observed results are inconsistent with a null hypothesis that the systems are statistically identical.

Tables 1 and 2 show results from the test that is usually most discriminative, the Matched Pairs Sentence Segment Word Error test. This test examines segments within turns for which the two systems disagreed but which are bounded on either side by at least two words that both systems recognized correctly [5]. The table entries show which system had the higher performance statistic where a significant difference is found, or a "~" when no significant difference is found. Where a difference is found, the number of asterisks indicates the strongest significance level (*~95%, **~99%, or ***~99.9%) found to hold in the comparison.

Note that a significant difference was found for all system pairs for the Switchboard type data, and for all but one pair with the CallHome data. In only one case, that of AT&T and BBN, was the direction of the difference found different for the two corpora.

## 5. TRENDS

Figures 3 shows multi-year histories of best performance results on the NIST conversational speech evaluations, including the best WER results for the past five evaluations on Switchboard type data (Switchboard-1 or Switchboard-2) and the best WER results for the past four evaluations on CallHome English data. In both cases the 2000 evaluation produced large, indeed dramatic, apparent performance advances from the previous evaluation. For Switchboard the relative percentage decrease in WER is 47%; for CallHome it is 25%.

It may be noted that the 1995 and 1996 evaluations used conversations from Switchboard-1, while the 1997 and 1998 evaluations used conversations from Switchboard–2, and that there was a large word error rate decrease in 1996 but not in 1997 and 1998. The return to Switchboard-1 type conversations this year was certainly a major reason for the bigger decrease from 1998 in Switchboard than in CallHome word error rates. While all Switchboard speakers were assigned a topic, Switchboard-1 speakers almost always kept to the topic, but Switchboard-2 speakers most often did not. Less “serious” conversations

| MATCHED PAIRS SENTENCE SEGMENT WORD ERROR SIGNIFICANCE TESTS FOR SWITCHBOARD TYPE CONVERSATIONS | | | | |
|---|---|---|---|---|
| | **BBN** | **CU-HTK** | **MS-State** | **SRI** |
| **AT&T** | AT&T * | CU-HTK *** | AT&T *** | AT&T *** |
| **BBN** | | CU-HTK *** | BBN *** | BBN ** |
| **CU-HTK** | | | CU-HTK *** | CU-HTK *** |
| **MS-State** | | | | SRI *** |

**Table 1**: Significant differences found by the Matched Pairs Sentence test for the Switchboard type conversations. The *'s show the maximum significance level, * for 95%, ** for 99%, and *** for 99.5%.

| MATCHED PAIRS SENTENCE SEGMENT WORD ERROR SIGNIFICANCE TESTS FOR CALLHOME CONVERSATIONS | | | | |
|---|---|---|---|---|
| | **BBN** | **CU-HTK** | **MS-State** | **SRI** |
| **AT&T** | BBN *** | CU-HTK *** | AT&T *** | ~ |
| **BBN** | | CU-HTK *** | BBN *** | BBN * |
| **CU-HTK** | | | CU-HTK *** | CU-HTK *** |
| **MS-State** | | | | SRI *** |

**Table 2:** Significant differences found by the Matched Pairs Sentence test for the CallHome conversations. The *'s show the maximum significance level, * for 95%, ** for 99%, and *** for 99.5%.

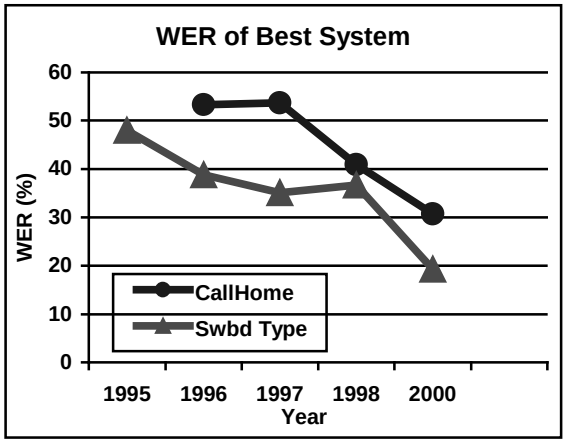


**Figure 3:** History of lowest word error rates obtained in NIST conversational speech evaluations on **S**witchboard and CallHome type conversations in English

with frequent changes in the subject of discourse are more difficult to automatically transcribe. The Switchboard-2 collection protocol presented the topic as more of a suggestion than as a specific directive, as in Switchboard-1. It may also be relevant that the Switchboard-2 speakers were generally younger and in college.

Another factor that has made the Switchboard-2 test sets more difficult for evaluation participants is that there has been considerably more transcribed Switchboard-1 data available for training than transcribed Switchboard-2 data. A factor clearly making this year's Switchboard test set easier was the inclusion in the training data of some speech from most of the test speakers. The performance impact of this appears to be fairly minor, however. Appendix 1 discusses this and other factors that may have contributed to making this year's test sets easier than those used in 1998.

Dragon Systems was a participant in previous NIST conversational speech recognition evaluations, but not a

participant in 2000. After the evaluation was over, however, they were kind enough to run their system from the 1998 evaluation on the 2000 test sets. Figure 4 shows the word error rates for this system on both the 1998 and 2000 test sets. Note that this Dragon system is not to be regarded as a competitive system for the 2000 evaluation.

This one "fixed" system showed a 39% relative decrease in word error rate on the Switchboard type data from 1998 to 2000, and a 16% relative decrease on the corresponding CallHome data sets. This very much suggests that the true gains due to system performance improvements by other sites in the 2000 evaluation were far smaller than those suggested by Figure 3. Most of the apparent gains were due to easier test sets in 2000 than in 1998. Appendix 1 discuss some of the factors that may have caused this year's test sets to be so much easier.

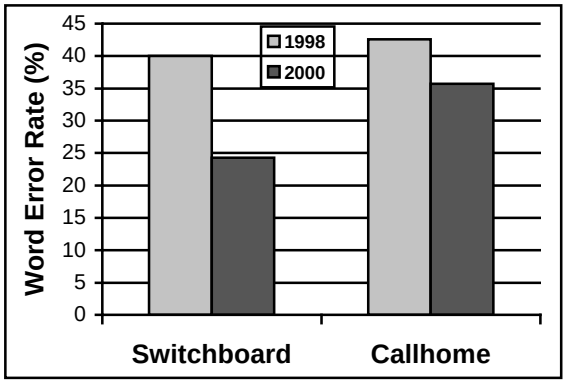


**Figure 4**: Word error rates for the Dragon 1998 system on the 1998 and 2000 test sets

## 6. MANDARIN EVALUATION

BBN Systems was the only site that chose to participate in the Mandarin evaluation this year. This evaluation used as test data twenty conversations from the CallHome Mandarin Corpus that had previously been set aside for evaluation purposes. Because of the nature of the language and its written representation, the metric used was Character Error Rate (CER) in accordance with the practice followed in previous Mandarin evaluations. BBN's submitted results produced a character error rate of 57.1%.

Although the CER is not directly comparable to the WER of the English evaluations, estimates made in previous Mandarin evaluations suggest that the character error rate is only moderately higher than a reasonably defined word error rate for the same conversation transcriptions. On this basis one may conclude that Mandarin recognition performance, as in previous evaluations, is considerably inferior to English performance. One key reason for this is the smaller amount of conversational telephone speech data available for acoustic training in Mandarin.

Figure 5 shows the CER's of the best performing Mandarin system in recent evaluations. All of these evaluations used as test data twenty Mandarin CallHome conversations. It is somewhat disappointing that the past trend toward lower error rates did not continue with this year's.

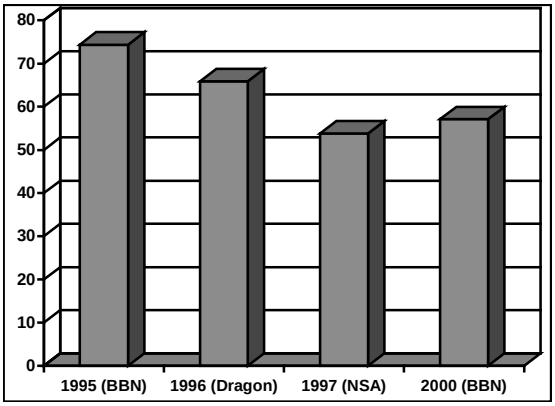


**Figure 5**: Character error rates of the best performing evaluation system in NIST Mandarin conversational speech evaluations 1995-2000

## 7. SUMMARY

The 2000 NIST evaluation of conversational telephone speech produced the lowest word error rates yet recorded on both Switchboard and CallHome type test sets. While the greater part of the word error rate decreases compared with the previous evaluation in 1998 are attributable to the test sets being easier, it is also clear that the leading 2000 systems offered significant real performance improvements over those in previous evaluations.

Another NIST conversational speech evaluation is anticipated for 2001. One question that will need to be decided is whether it should include Switchboard-1 or Switchboard-2 type conversations. This 2000 evaluation has served to highlight the real differences between these.

There is a need for continuing consideration of the issue of how to make evaluation test sets as comparable as possible between evaluations. There is also clearly a continuing need for high quality "fixed" recognition systems to calibrate the differences in test set difficulty.

## APPENDIX 1: TEST SET DIFFICULTY FACTORS

We consider in this appendix a variety of factors that may have contributed to making the 2000 evaluation test sets easier for automatic conversational speech recognizers than the test sets in the 1998 evaluation. We concentrate mainly on the Switchboard test sets, which had the greater performance differences between the two evaluations.

### Test Data Included in Training

The forty speakers in the 2000 Switchboard test set each appeared in between 0 and 8 of the original Switchboard conversations available for training for the evaluation. Figure 6 shows the range of performance of the evaluation systems for these speakers as a function of the number of training conversations in which they appear. There is a very mild trend toward decreasing word error rate as the number of training conversations increases. This is clearly,

however, a minor factor in performance. While having evaluation test speakers appear in the training data was certainly regrettable, it did not have a major effect on the performance results.

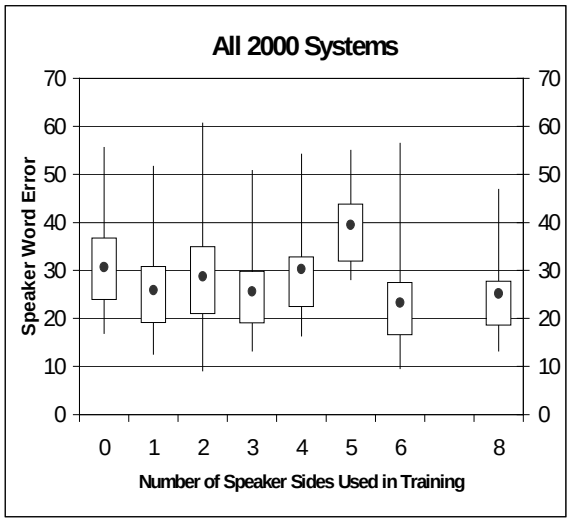


**Figure 6**: Ranges of word error rates of 2000 Switchboard conversation sides as a function of the number of training sides containing the speaker

## Disfluencies

The disfluencies common in conversational speech are generally viewed as one of the major sources of difficulty for automatic recognition in this environment. These disfluencies include word fragments and hesitation sounds, and while these items themselves are optionally deletable for scoring purposes, i.e. are not scored as an error if missing, they are likely to hinder recognition of surrounding words because of the language model employed. Figure 7 shows the disfluency rates for the 1998 and 2000 Switchboard type test sets. This difference in disfluency rates is apparently highly significant ($p < .0001$) and is a possible cause of the decreased error rate for 2000. Some caution may be appropriate, however, as the transcription and annotation of the conversations of the two test sets were done at different times by different organizations, and this could be partially responsible for the apparent disfluency differences.

## Signal-to-Noise Ratio

Signal-to-noise ratio (SNR) is certainly a factor that may affect performance, though all of the data sets in both the 1998 and 2000 evaluations are fairly clean. NIST's software to estimate segmental peak SNR has been adapted to handle mu-law data and to make use of the time-mark information available in the reference transcriptions of conversations. Figure 8 shows the ranges of SNR values found for the 1998 and 2000 test sets. More details on the test sets procedures may be obtained from William Fisher.

The SNR ranges found suggest that quieter data in 2000 could be a factor in the improved performance over 1998.

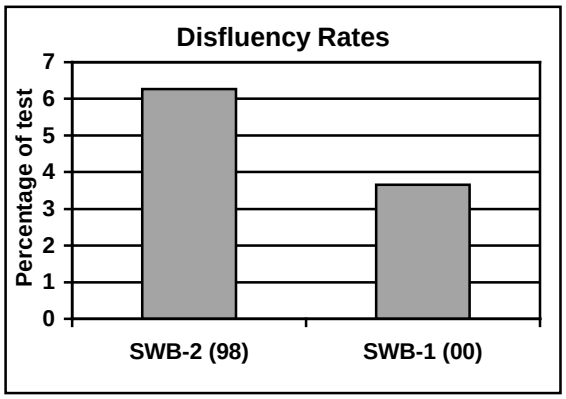


**Figure 7:** Disfluency rates of Switchboard 1998 and 2000 test sets

The larger difference between the two years occurs with the Switchboard conversations. This is presumably because for Switchboard-2 the initiators were required to use unique phone lines for each call, and they sometimes used public phones in relatively noisy environments. It is not clear why the CallHome calls in 2000 should have been quieter than those from the same corpus in 1998. But the differences are not great.

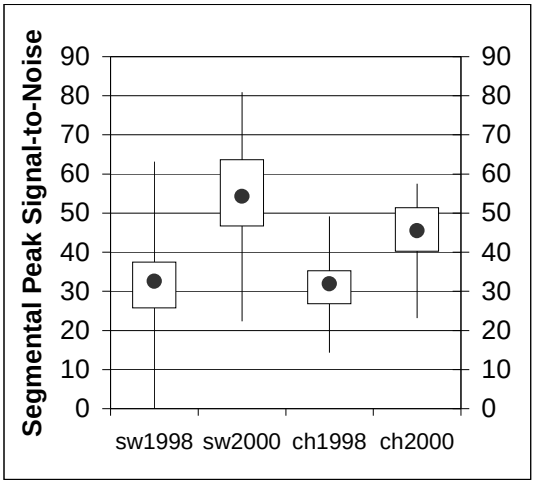


**Figure 8**: Segmental peak SNR ratios for the 1998 and 2000 Switchboard and CallHome test sets

## Segment Durations

We also examined the effect of segment (i.e., turn) durations on performance for the Switchboard test sets. Figure 9 is a scatterplot of word error rates for the Dragon '98 system as a function of average segment durations, measured in words, of conversation sides.

It is clear from Figure 9 that the turn durations are frequently longer for the Switchboard-1 type data used in 2000 than for the Switchboard-2 data used in 1998. This is probably because of the more topical nature of the conversations in 2000, as discussed previously.

It is less clear how much this average duration difference

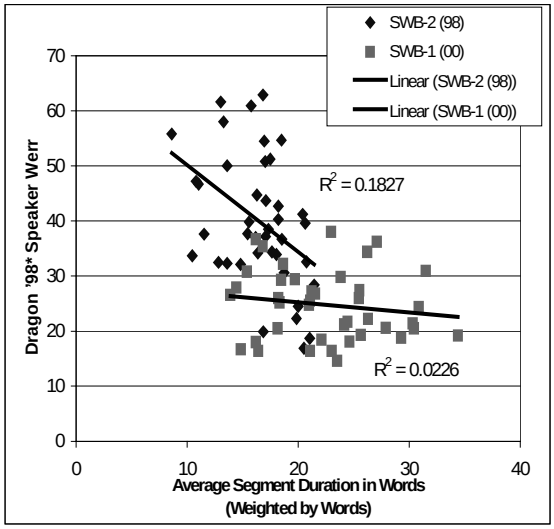


**Figure 9**: Scatterplot of average conversation side segment durations and word error rates for the Switchboard 1998 and 2000 test sets with the Dragon'98 recognizer

affects performance. The word error rates for the Switchboard-2 conversation sides from 1998 appear rather more sensitive to average duration (as shown by the steeper negative slope of the fitted curve) than do the 2000 conversation sides. It may also be that the effect of longer average duration on performance fades away for average durations exceeding about 20 words.

## Perplexity

The perplexity of a test set relative to a language model is known to be a factor likely to affect speech recognition performance. Figure 10 (courtesy of Roni Rosenfeld of Carnegie Mellon University) shows the perplexities of the Switchboard test sets from 1996 to 2000 for a standard language model derived and from the available transcribed Switchboard-1 and Switchboard-2 data. Note that the 2000 test set has a 10-12% lower perplexity than the 1998 test set. Note as well the lower perplexities of the 1996 Switchboard-1 test sets, compared with those from Switchboard-2. It is not clear to what extent this reflects the inherent differences between the two Switchboard Corpora, and to what extent it may be due to the much greater amount of transcribed training data available from Switchboard-1.

## Segmentation Procedures

One difference between the 1998 and 2000 Switchboard type test set that has not been carefully examined is the different segmentation procedures used to define the turns. The ISIP-provided software used this year is likely to become the standard for use in newly collected conversational data. While it should not result in performance particularly different from that when the previous BBN segmentation procedure was used, it would be desirable to conduct a specific comparison of recognition performance with the two segmentation procedures using a common recognizer and a common set

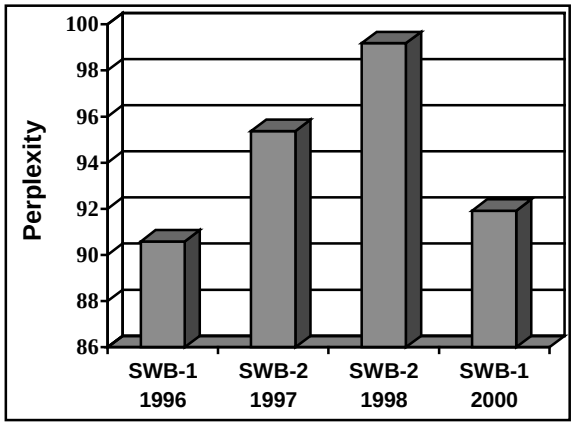


**Figure 10**: Perplexities of Switchboard test sets 1996-2000

of conversations.

## Summary

There thus appear to have been various factors related to this year's Switchboard test data being easier than that in previous evaluations. The most basic would appear to be the different nature of the Switchboard-1 and Switchboard-2 Corpora. Switchboard-2 type conversations are clearly the greater challenge for automatic speech recognizers, but the Switchboard-1 type are challenging enough given the current state of the technology, and some would argue that that type of conversation offers the more interesting and useful application.

It is less clear why this year's CallHome test set was easier than that of the previous evaluation, though this was a smaller difference. One factor may be that after the 1997 evaluation there was an effort to weed out CallHome English conversations dominated by speech that would not be regarded as "standard American English" speech. This should, however, have affected the 1998 evaluation as well. It may be that the differences in difficulty between the 1998 and 2000 CallHome test sets are within the margin to be expected by chance for sets of this size. Such differences have occurred in other evaluations.

All of this further suggests the value in having available a standard, relatively low error rate "fixed" recognizer which may be used to accurately assess test set differences. Dragon's assistance this year with their 1998 system was invaluable as a diagnostic tool.

# APPENDIX 2: SNR

One possible factor making this year's test sets easier would be higher signal-to-noise ratios. This appendix analyses the possible significance of this factor.

Because blind estimation of the signal-to-noise ratio (SNR) of CODEC-encoded speech has not been very successful in the past, we developed a new method for estimating it.

In this method, separation of the noise and speech signals is

accomplished by force-aligning a reference transcription to the acoustic signal (instead of through use of an energy-based segmenter such as the QA software's "speech" program). Given such a time-marked alignment, our current algorithm classifies segments stretching from a sentence-beginning to the next sentence-ending as a "speech" segment (actually speech plus noise), and calling segments from a sentence-ending to the next sentence-beginning "noise" segments. This simple scheme has two further embellishments: 1) sentence-initial and -final /sil/ phones are considered part of the noise, rather than speech; and 2) noise segments longer than 15 seconds are not used, since it's likely that they may contain untranscribed speech.

Given such "noise" and "speech" segments, we calculated a measure similar to SNR using a variant of the program "segsnr" that was created at NIST several years ago for the Speech Quality Assurance software project. This accepts a list of speech segments, usually from a speech detector, and assumes the rest of the file is noise. It estimates the power of the speech segments as the 95th percentile of the power histogram of their frames. The power of the noise segments is estimated by computing their average y**2 value. And finally, the peak speech to noise ratio (SPNR) is estimated as the ratio of the two expressed in decibels re 1.0. Our change was just to use an explicit marking of the noise segments, instead of assuming that all of the file not classified as speech is noise. This was necessary because some parts of the speech files were left untranscribed.

Estimating energy of noise from a signal that has been companded with a CODEC chip is complicated by the fact that low-energy noise is usually quantized to zero. Indeed, the most prominent mode of the Switchboard noise power histogram is at -20.0 dB. We chose to ignore frames whose estimated power is less than 2.0 dB. This results in the loss of 22.9% of the total of 118.075 frames, almost all of them -20 dB frames from the Switchboard 1998 data. It should be recognized that this could be affecting the results presented here.

In order to work with a unit that contains both speech and noise, we aggregate these basic segments into conversation-side or "speaker" units.

A preliminary look at a scatter plot of WER vs. SPNR for all speakers, shown below as Figure 11, suggested a significant correlation in the broad middle range of SPNR, with little or none at the extreme high and low ends. This figure shows all data from both Callhome and Switchboard, from all systems, with only pathological points having a speech rate greater than 700 words/minute being excluded. The scatter is large, due to the uncontrolled effects of many other variables. The polynomial of degree three that was fit to the data appears to show a smooth downward slope in the lower to middle range of the data. We therefore limited our further study to speakers with mean SPNR in the range of 25-55db. The correlation coefficient of WER and SPNR in this range is -0.218, which is statistically significant (1-tail, n=153, $p < .025$) although not large. To test if this method is an improvement over the "segsnr" method in our standard QA software using the "speech" segmenter, we computed the correlation coefficient using that method over the same speaker sides; it came out to be only -0.0978,

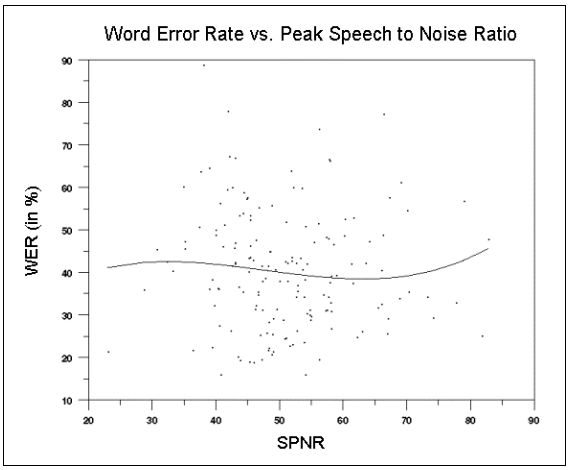


**Figure 11**: 1998 & 2000 Speaker mean word error rate vs. segmental SPNR.

clearly less, and not significant. In this range, the best-fitting linear function is: WER = 63.77 - 0.54*SPNR

In terms of a first-order functional model, every increase in SPNR of 2 dB in this range increases the WER on average by a little over 1.0.

Again restricting our view to the 25-55 dB broad middle range where SPNR seems to count, Table 3 presents the speaker-averaged mean values of SPNR for the different partitions of the evaluation data in which we are interested.

| | 1998 | 2000 | Delta |
|---|---|---|---|
| Switchboard | 39.92 | 43.58 | 3.66 |
| CallHome | 44.09 | 45.42 | 1.33 |
| Delta | 4.17 | 1.84 | |

**Table 3**: Speaker-averaged mean values of peak speech to noise ratios

Testing significance with a 2-tailed Mann-Whitney test, the significant ($p < .001$) generalizations about these differences are:

- For 1998, Switchboard had a lower SPNR than Call Home, accounting for an increase in WER of about 2.25.
- For 2000, Switchboard had a lower SPNR than in 1998, accounting for a decrease in WER of about 2.0.

The other differences, although consistent with these, cannot be shown to be statistically significant

## ACKNOWLEDGEMENTS

NIST acknowledges the assistance of Barbara Peskin and others at Dragon Systems in running the 1998 Dragon speech recognizer on the 2000 data sets. It should again be emphasized that Dragon was in no way an entrant in the 2000 evaluation.

NIST also acknowledges the assistance of Roni Rosenfeld of Carnegie Mellon University in generating the test set

perplexity estimates presented above.

## NOTICE

The views expressed in this paper are those of the authors. The test results are for local, system-developer implemented tests. NIST's role was one that involved working with the LDC in processing LDC-provided training and test speech data and reference transcriptions, developing and implementing scoring software, and uniformly scoring and tabulating results. The views of the authors, and these results, are not to be construed or represented as endorsements of any systems, or as official findings on the part of NIST, DARPA, or the U.S. Government.

## REFERENCES


1. Greenberg, S., Chang, S., and Hollenback, J., "An Introduction to the Diagnostic Evaluation of Switchboard-Corpus Automatic Speech Recognition Systems", Proceedings 2000

2. J. Fiscus, "A Post-Processing System To Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", Proceedings, 1997 IEEE Workshop on Automatic Speech Recognition and Understanding

3. The 2000 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone, URL - http://www.itl.nist.gov/iaui/894.01/tests/ctr/h5_2000/h5-2000-v1.3.htm

4. "The Confidence Measure" in The 2000 NIST Evaluation Plan for Recognition of Conversational Speech over the Telephone, URL - http://www.itl.nist.gov/iaui/894.01/tests/ctr/h5_2000/h5-2000-v1.3.htm

5. Matched Pairs Sentence-Segment Word Error (MAPSSWE) Test, URL - http://www.itl.nist.gov/iaui/894.01/tests/sigtests/mapsswe.htm