PERFORMANCE OF SRI'S DECIPHER™ SPEECH RECOGNITION SYSTEM ON DARPA'S CSR TASK

Hy Murveit, John Butzberger, and Mitch Weintraub

SRI International
Speech Research and Technology Program
Menlo Park, CA, 94025

1. ABSTRACT

SRI has ported its DECIPHER™ speech recognition system from DARPA's ATIS domain to DARPA's CSR domain (read and spontaneous *Wall Street Journal* speech). This paper describes what needed to be done to port DECIPHER™, and reports experiments performed with the CSR task.

The system was evaluated on the speaker-independent (SI) portion of DARPA's February 1992 "Dry-Run" WSJ0 test and achieved 17.1% word error without verbalized punctuation (NVP) and 16.6% error with verbalized punctuation (VP). In addition, we increased the amount of training data and reduced the VP error rate to 12.9%. This SI error rate (with a larger amount of training data) equalled the best 600-training-sentence speaker-dependent error rate reported for the February CSR evaluation. Finally, the system was evaluated on the VP data using microphones unknown to the system instead of the training-set's Sennheiser microphone and the error rate only increased to 26.0%.

2. DECIPHER™

The SRI has developed the DECIPHER™ system, an HMM-based speaker-independent, continuous-speech recognition system. Several of DECIPHER™'s attributes are discussed in the references (Butzberger et al., [1]; Murveit et al., [2]). Until recently, DECIPHER™'s application has been limited to DARPA's resource management task (Pallet, [3]; Price et al., [4]), DARPA's ATIS task (Price, [5]), the Texas Instruments continuous-digit recognition task (Leonard, [6]), and other small vocabulary recognition tasks. This paper describes the application of DECIPHER™ to the task of recognizing words from a large-vocabulary corpus composed of primarily read-speech.

3. THE CSR TASK

Doddington [7] gives a detailed description of DARPA's CSR task and corpus. Briefly, the CSR corpus* is composed of recordings of speakers reading passages from the *Wall Street Journal* newspaper. The corpus is divided in many ways; it includes speaker-dependent vs. speaker independent sections and sentences where the users were asked to verbalize the punctuation (VP) vs. those where they were asked not to verbalize the punctuation (NVP). There are also a small number of recordings of spontaneous speech that can be used in development and evaluation.

The corpus and associated development and evaluation materials were designed so that speech recognition systems may be evaluated in an open-vocabulary mode (none of the words used in evaluation are known in advance by the speech recognition system) or in a closed vocabulary mode (all the words in the test sets are given in advance). There are suggested 5,000-word and 20,000-word open- and closed-vocabulary language models that may be used for development and evaluation. This paper discusses a preliminary evaluation of SRI's DECIPHER™ system using read speech from the 5000-word closed-vocabulary tasks with verbalized and nonverbalized punctuation.

4. PORTING DECIPHER™ TO THE CSR TASK

Several types of data are needed to port DECIPHER™ to a new domain:

- A target vocabulary list
- A target language model
- Task-specific training data (optional)
- Pronunciations for all the words in the target vocabulary (mandatory) and for all the words in the training data (optional)
- A backend which converts recognition output to actions in the domain (not applicable to the CSR task).

*The current CSR corpus, designated WSJ0 is a pilot for a large corpus to be collected in the future.

4.1. CSR Vocabulary Lists and Language Models

Doug Paul at Lincoln Laboratories provided us with baseline vocabularies and language models for use in the February 1992 CSR evaluation. This included vocabularies for the closed vocabulary 5,000 and 20,000-word tasks as well as backed-off bigram language models for these tasks. Since we used backed-off bigrams for our ATIS system, it was straightforward to use the Lincoln language models as part of the DECIPHER™-CSR system.

4.2. CSR Pronunciations

SRI maintains a list of words and pronunciations that have associated probabilities automatically estimated (Cohen et al., [8]). However, a significant number of words in the speaker-independent CSR training, development, and (closed vocabulary) test data were outside this list. Because of the tight schedule for the CSR evaluation, SRI looked to Dragon Systems which generously provided SRI and other DARPA contractors with limited use of a pronunciation table for all the words in the CSR task. SRI combined its internal lexicon with portions of the Dragon pronunciation list to generate a pronunciation table for the DECIPHER™-CSR system.

4.3. CSR Training Data

The National Institute of Standards and Technology provided to SRI several CDROMS containing training, development, and evaluation data for the February 1992 DARPA CSR evaluation. The data were recorded at SRI, MIT, and TI. The baseline training conditions for the speaker-independent CSR task include 7240 sentences from 84 speakers, 3,586 sentences from 42 men and 3,654 sentences from 42 women.

5. PRELIMINARY CSR PERFORMANCE

5.1. Development Data

We have partitioned the speaker-independent CSR development data into four portions for the purpose of this study. Each set contains 100 sentences. The respective sets are male and female speakers using verbalized and nonverbalized punctuation. There are 6 male speakers and 4 female speakers in the SI WSJ0 development data.

The next section shows word recognition performance on this development set using 5,000-word, closed-vocabulary language models with verbalized and nonverbalized bigram grammars. The perplexity of the verbalized punctuation sentences in the development set is 90.

5.2. Results for a Simplified System

Our strategy was to implement a system as quickly as possible. Thus we initially implemented a system using four vector-quantized speech features with no cross-word acoustic modeling. Performance of the system on our development set is described in the tables below.

Table 1: Simple Recognizer

Speaker	Verbalized Punctuation %word err	Non Verbalized Punctuation %word err
050	10.0	11.8
053	14.0	17.6
420	14.7	18.1
421	11.9	17.9
051	21.1	18.8
052	20.7	20.2
22g	15.4	19.6
22h	20.8	13.0
422	57.9	40.4
423	15.0	24.6
Average	20.1	20.2

The female speakers are those above the bold line in Table 1. Recognition speed on a Sun Sparcstation-2 was approximately 40 times slower than real time (over 4 minutes/sentence) using a beam search and no fast match (our standard smaller-vocabulary algorithm), although it was dominated by paging time.

A brief analysis of Speaker 422 shows that he speaks much faster than the other speakers which may contribute to the high error rate for his speech.

5.3. Full DECIPHER™-CSR Performance

We then tested a larger DECIPHER™ system on our VP development set. That is, the previous system was extended to model some cross-word acoustics, increased from four to

six spectral features (second derivatives of cepstra and energy were added) and a tied-mixture hidden Markov model (HMM) replaced the vector-quantized HMM above. This resulted in a modest improvement as shown in the Table 2.

Table 2: Full Recognizer

Speaker	Verbalized Punctuation %word err
050	11.1
053	11.7
420	13.7
421	11.0
051	20.0
052	14.2
22g	15.7
22h	14.9
422	48.3
423	13.0
Average	17.4

6. DRY-RUN EVALUATION

Subsequent to the system development, above, we evaluated the "full recognizer" system on the February 1991 Dry-Run evaluation materials for speaker-independent systems. We achieved word error rates of 17.1% without VP and 16.6% error rates with VP as measured by NIST.*

Table 3: Dry-Run Evaluation Results

Speaker	Non Verbalized Punctuation %word err	Verbalized Punctuation %word err
427	9.4	9.0
425	20.1	15.1
z00	14.4	16.7
063	24.5	17.8
426	10.2	10.8
060	17.0	22.9
061	12.3	13.6
22k	25.3	17.6
221	17.8	12.4
424	20.0	18.4
Average	17.1	15.4

7. OTHER MICROPHONE RESULTS

The WSJ0 corpus was collected using two microphones simultaneously recording the talker. One was a Sennheiser HMD-410 and the other was chosen randomly for each speaker from among a large group of microphones. Such

*The NIST error rates differ slightly (insignificantly) from our own measures (17.1% and 16.6%), however, to be consistent with the other error rates reported in this paper, we are using our internally measured error rates in the tables.

dual recordings are available for the training, development, and evaluation materials.

We chose to evaluate our full system on the "other-microphone" data without using other-microphone training data. The error rate increased only 62.3% when evaluating with other-microphone recordings vs. the Sennheiser recordings.

In these tests, we configured our system exactly as for the standard microphone evaluation, except that we used SRI's noise-robust front end (Erell and Weintraub, [9,10]; Murveit, et al., [11]) as the signal processing component.

Table 4 summarizes the "other-microphone" evaluation results. Speaker 424's performance, where the error rate increases 208.2% (from 18.4% to 56.7%) when using a Shure SM91 microphone is a problem for our system. However, the microphone is not the sole source of the problem, since the performance of Speaker 427, with the same microphone, is only degraded 18.9% (from 9.0 to 10.7%). We suspect that the problem is due to a loud buzz in the recordings that is absent from the recordings of other speakers.

8. EXTRA TRAINING DATA

We suspected that the set of training data specified as the baseline for the February 1992 Dry Run Evaluation was insufficient to adequately estimate the parameters of the DECIPHER™ system. The baseline SI training condition contains approximately 7,240 from 84 speakers (half42 male, 42 female).

We used the SI and SD training and development data to train the system to see if performance could be improved with extra data. However, to save time, we used only speech from male speakers to train and test the system. Thus, the training data for the male system was increased from 3586 sentences (42 male speakers) to 9109 sentences (53 male speakers).* The extra training data reduced the error rate by approximately 20% as shown in Table 5.

*The number of speakers did not increase substantially since the bulk of the extra training data was taken from the speaker-dependent portion of the corpus.

Table 4: Verbalized Punctuation Evaluation Results Using "Other Microphones"

Speaker	Microphone	%word error "other mic"	%word error Sennheiser mic	%degradation
427	Shure SM91 desktop	10.7	9.0	18.9
425	Radio Shack Highball	21.4	15.1	41.8
z00	Crown PCC160 desktop	24.9	16.7	49.1
063	Crown PCC160 desktop	29.4	17.8	65.2
426	ATT720 telephone over local phone lines	12.1	10.8	12.0
060	Crown PZM desktop	30.5	22.9	33.2
061	Sony ECM-50PS lavaliere	18.8	13.6	38.2
22k	Sony ECM-55 lavaliere	25.3	17.6	43.8
221	Crown PCC160 desktop	22.8	12.4	83.9
424	Shure SM91 desktop	56.7	18.4	208.2
Average		25.0	15.4	62.3

Table 5: Evaluation Male Speakers with Extra Training Data

Speaker	Baseline Training	Larger-Set Training
060	22.6	15.5
061	13.6	8.2
22k	17.6	16.8
221	12.4	11.3
424	18.4	15.7
426	10.8	9.8
Average	15.8	12.9

Interestingly, this reduced error rate equalled that for speaker-dependent systems trained with 600 sentences per speaker and tested with the same language model used here. However, speaker-dependent systems trained on 2000+ sentences per speaker did perform significantly better than this system.

9. SUMMARY

This is a preliminary report demonstrating that the DECIPHER™ speech recognition system was ported from a 1,000-word task (ATIS) to a large vocabulary (5,000-word) task (DARPA's CSR task). We have achieved word error rates between of 16.6% and 17.1% as measured by NIST on DARPA's February 1992 Dry-Run WSJ0 evaluation where no test words were outside the prescribed vocabulary. We evaluated using alternate microphone data and found that the error rate increased only by 62%. Finally, by increasing the amount of training data, we were able to achieve an error rate that matched the error rates reported for this task from 600 sentence/speaker speaker-dependent systems. This could not have been done without substantial support from the rest of the DARPA community in the form of speech data, pronunciation tables, and language models.

ACKNOWLEDGEMENTS

We gratefully acknowledge support for this work from DARPA through Office of Naval Research Contract N00014-90-C-0085. The Government has certain rights in this material. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the government funding agencies.

We would like to that Doug Paul at Lincoln Laboratories for providing us with the Bigram language models used in this study, and Dragon Systems for providing us with the Dragon pronunciations described above. We would also like to thank the many people at various DARPA sites involved in specifying, collecting, and transcribing the speech corpus used to train, develop, and evaluate the system described.

REFERENCES

1. Butzberger, J., H. Murveit, E. Shriberg, and P. Price, "Modeling Spontaneous Speech Effects in Large Vocabulary Speech Recognition," DARPA SLS Workshop Proceedings, Feb 1992.

2. Murveit, H., J. Butzberger, and M. Weintraub, "Speech Recognition in SRI's Resource Management and ATIS Systems," DARPA SLS Workshop, February 1991, pp. 94-100.

3. Pallet, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *IEEE ICASSP 1989*, pp. 536-539.

4. Price, P., W.M. Fisher, J. Bernstein, and D.S. Pallet, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE ICASSP 1988*, pp. 651-654.

5. Price, P., "Evaluation of SLS: the ATIS Domain," DARPA SLS Workshop, June 1990, pp. 91-95.

6. Leonard, R.G., "A Database for Speaker-Independent Digit Recognition," *IEEE ICASSP 1984*, p. 42.11

7. Doddington, G., "CSR Corpus Development," DARPA SLS Workshop, Feb 1992.

8. Cohen, M., H. Murveit, J. Bernstein, P. Price, and M. Weintraub, "The DECIPHER™ Speech Recognition System," *IEEE ICASSP-90*.

9. Erell, A., and M. Weintraub, "Spectral Estimation for Noise Robust Speech Recognition," DARPA SLS Workshop October 89, pp. 319-324.

10. Erell, A., and M. Weintraub, "Recognition of Noisy Speech: Using Minimum-Mean Log-Spectral Distance Estimation," DARPA SLS Workshop, June 1990, pp. 341-345.

11. Murveit, H., J. Butzberger, and M. Weintraub, "Reduced Channel Dependence for Speech Recognition", DARPA SLS Workshop Proceedings, February 1992.