# Training Set Issues in SRI's DECIPHER Speech Recognition System

**Hy Murveit, Mitch Weintraub, Mike Cohen**

Speech Research Program
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

## Abstract

SRI has developed the DECIPHER system, a hidden Markov model (HMM) based continuous speech recognition system typically used in a speaker-independent manner. Initially we review the DECIPHER system, then we show that DECIPHER's speaker-independent performance improved by 20% when the standard 3990-sentence speaker-independent test set was augmented with training data from the 7200-sentence resource management speaker-dependent training sentences. We show a further improvement of over 20% when a version of corrective training was implemented. Finally we show improvement using parallel male- and female-trained models in DECIPHER. The word-error rate when all three improvements were combined was 3.7% on DARPA's February 1989 speaker-independent test set using the standard perplexity 60 wordpair grammar.

## System Description

### Front End Analysis

Decipher uses a FFT-based Mel-cepstra front end. Twenty-five FFT-Mel filters spanning 100 to 6400 hz are used to derive 12 Mel-cepstra coefficients every 10-ms frame. Four features are derived every frame from this cepstra sequence. They are:

- Vector-quantized energy-normalized Mel-cepstra
- Vector-quantized smoothed 40-ms time derivatives of the Mel-cepstra
- Energy
- Smoothed 40-ms energy differences

We use 256-word speaker-independent codebooks to vector-quantize the Mel-cepstra and the Mel-cepstral differences. The resulting four-feature-per-frame vector is used as input to the DECIPHER HMM-based speech recognition system.

### Pronunciation Models

DECIPHER uses pronunciation models generated by applying a phonological rule set to word baseforms. The technique used to generate the rules are described in Murveit89 and Cohen90. These generate approximately 40 pronunciations per word as measured on the DARPA resource management vocabulary. Speaker-independent pronunciation probabilities are then estimated using these bushy word networks and the forward-backward algorithm in DECIPHER. The networks are then pruned so that only the likely pronunciations remain--typically about four pronunciations per word for the resource management task.

This modeling of pronunciation is one of the ways that DECIPHER is distinguished from other HMM-based systems. We have shown in Cohen90 that this modeling improves system performance.

### Acoustic Modeling

DECIPHER builds and trains word models by using context-based phone models arranged according to the pronunciation networks for the word being modeled. Models used include unique-phone-in-word, phone-in-word, triphone, biphone, and generalized-phone forms of biphones and triphones, as well as context-independent models. Similar contexts are automatically smoothed together, if they do not adequately model the training data, according to a deleted-estimation interpolation algorithm developed at SRI (similar to Jelinek80). The acoustic models reflect both inter-word and across-word coarticulatory effects.

Training proceeds as follows:

- Initially, context-independent boot models are estimated from hand-labeled portions of the training part of the TIMIT database.
- The boot models are used as input for a 2-iteration context-independent model training run, where context-independent models are refined and pronunciation probabilities are calculated using the large 40-pronunciation word networks. As stated above, these large networks are then pruned to about four pronunciations per word.

- Context-dependent models are then estimated from a second 2-iteration forward-backward run, which uses the context-independent models and the pruned networks as input.

### System Evaluation

DECIPHER has been evaluated on the speaker-independent continuous-speech DARPA resource management test sets [Price88] [Pallet89]. DECIPHER was evaluated on the November 1989 test set (evaluated by SRI in March 1990) and had 6% word error on the perplexity 60 task. This performance was equal to the best previously reported error rate for that condition. We recently evaluated on the June 1990 task, and achieved 6.5% word error for a system trained on 3990 sentences and 4.8% word error using 11,190 training sentences.

Since the October 1989 evaluation, DECIPHER's performance has improved in three ways:

- We noted when using that the standard 3990-sentence resource management training set, that many of DECIPHER's probability distributions were poorly estimated. Therefore, we evaluated DECIPHER with several different amounts of training data. The largest training set we used, an 11,190-sentence resource management training set, improved the word error rate by about 20%.
- We implemented a modified version of IBM's corrective training algorithm, additionally improving the word error rate by about 20%.
- We separated the male and female training data, estimated different HMM output distributions for each sex. This also improved word accuracy by 20%.

These improvements are described in more detail below.

## Effects of Training Data

In a recent study, we discovered that DECIPHER's word error rate on its training set using the perplexity 60 grammar was very low (0.7% over the 3990 resource management sentences). Since the test-set error rate for that system was about 7%, we concluded that the system would profit from more training data. To test this, we evaluated the system with four databases easily available to us as is shown in Table 1. There *SI* refers to the 3990-sentence speaker-independent portion of the resource management (RM) database--109 speakers, 30 or 40 sentences each, *SD* refers to the speaker-dependent portion of that database--12 speakers, 600 sentences each, and *TIMIT* refers to the training portion of the TIMIT database--420 speakers, 8 sentences each. Note that all SI and SD sentences are related to the resource management task, while TIMIT's sentences are not related to that task. All systems were tested using a continuous-speech, speaker-independent condition with the perplexity 60 resource management grammar testing on DARPA's 300-sentence February 1989 speaker-independent test set.

| Training data | Sentences | Word error |
|---|---|---|
| SD | 7200 | 7.3 |
| SI | 3990 | 6.7 |
| SI+TIMIT | 7350 | 5.8 |
| SI+SD | 11190 | 5.3 |

**Table 1.**
**Word Error as a Function of Training Set**

Table 1 shows that performance improved as data increased, even when adding the out-of-task TIMIT data. The only exception was that training with 3990 sentences from 100 talkers was slightly better than 7200 sentences from 12 talkers. This is to be expected in a speaker-independent system. This last result is consistent with the findings in Kubala90 that showed that there was not a big performance drop when the number of speakers was drastically reduced (from 109 to 12) in speaker-independent systems. It is likely that more training data would continue to improve performance on this task; however, we believe that a more sensible study would be to focus on how large training sets could improve performance across tasks and vocabularies. (See, for instance, Hon90.)

## Separating Male and Female Models

We experimented with maintaining sex consistency in DECIPHER's hypotheses by partitioning male and female training data and using parallel recognition systems as in Bush87. Two *subrecognizers* are run in parallel on unknown speech and the hypothesis from either recognizer with the highest probability is used. The disadvantage of this approach is that it makes inefficient use of training data. That is, in the best scenario the male models are trained from only half of the training data and the female models use only half. This is inefficient because even though there may be a fundamental difference between the two types of speech, they still have many things in common and could profit from the others' training data if used properly.

It is no wonder, then, that this approach has been successful in digit recognition systems with an abundance of training data for each parameter to be estimated, but has not significantly improved performance in large-vocabulary systems with a relatively small amount of training data [Paul89]. To validate the idea of sex consistency, we trained male-only and female-only versions of the DECIPHER speech recognition system using the 11190-sentence SI+SD training set to make sure the data partitions had enough data. We produced SI+SD

subsets with 4160 female and 7030 male sentences. These systems were tested on the DARPA February 1989 speaker-independent test set using the DARPA word-pair grammar (perplexity 60) and are compared below to a similar recognition system trained on all 11190 sentences.

| | Standard | Male/Female |
|---|---|---|
| Male speakers | 5.5 | 4.6 |
| Female speakers | 4.9 | 4.0 |
| All speakers | 5.3 | 4.3 |

**Table 2. Speaker-Independent %Word Error for Male/Female Parallel Recognizers (February 1989 SI Test Set)**

The results in Table 2 show a 19% reduction in the error rate when using sex-consistent recognition systems. This is a significant error rate reduction. A closer look at the system's performance showed that it correctly assigned the talker's sex in each of the 300 test sentences.

## Discriminative Techniques Currently in DECIPHER

We have implemented a type of corrective training [Bahl88, Lee89] in the DECIPHER system. Our implementation is similar to that described in Lee89 with the following exceptions or notes:

1. We use four partitions (rather than two) for our deleted estimation technique. In this way, the recognition systems used to generate alignments for corrective training are as similar as possible to the overall recognition system.
2. We do not alter the actual HMM counts for states, but rather scale the states' vector output probabilities by the ratio (#correct+#deletions-#insertions) divided by #correct. These counts are generated by frame alignments of the recognizer hypothesis and the correct sentence. This improves performance from 5.9% word error to 5.1% on the February 1989 test set using the standard SI training set--the uncorrected system has 6.7% word error. The reason for this improvement may be that adjusting the counts of a model affects other models (given our deleted interpolation estimation smoothing algorithms) that do not require correction. Scaling model probabilities only adjusts the models that require change.
3. We do not generate reinforcement errors. We plan to do so using an N-best algorithm to generate alternate hypotheses.
4. We can not iterate the algorithm until the N-best reinforcement is implemented, because the second iteration error rate on the sentences that had been corrected by the first iteration was under 0.3%.

Our implementation reduced the error rate on the February 1989 test set by 24% (6.7% to 5.1%) which is approximately the improvement gained by Lee89 and Bahl88.

Points 3 and 4 above are a concern, because they limit the efficiency with which this algorithm could use its already limited training data. To examine this, we performed the following two experiments. (1) We added a second pass of corrective training, using the speaker-dependent RM training sentences (SD). (2) We combined SD and the SI sentences, thereby using a larger overall training set, but continued to use one pass of corrective training. Table 3 shows that, not surprisingly, though

| System | Training | Word Error |
|---|---|---|
| no correction | SI | 6.7% |
| 1 pass correction | SI | 5.1% |
| add 2nd SD pass | SI | 4.6% |
| | | |
| no correction | SI+SD | 5.3% |
| 1 pass correction | SI+SD | 4.1% |

**Table 3. Corrective Training with Extra Data (Uses February 1989 RM Test Set)**

there was improvement when extra data were used as a second pass for the corrective training algorithm, it was better to use these data to simply augment the training data (4.6% versus 4.1% word error). It is also interesting to note that the improvement gained by corrective training with the 3990 SI sentences (6.7% to 5.1%, 24% fewer errors) was approximately equal to the improvement gained by applying corrective training to the larger 11190 SI+SD sentences (5.3% to 4.1%, 23% fewer errors). This leads us to believe that lack of training data is not more of a bottleneck for corrective training than it is for the system as a whole.

## Combining Corrective Training and Sex Consistency

We combined both sex consistency and corrective training and arrived at the improvement shown in Table 4. We didn't achieve the same 20% improvement as in the past, probably due to training data limitations.

Attempting the combined system with the standard 3990-sentence training set resulted in poor performance, primarily because the female models used to train

the corrective training partitions had only 870 sentences of training data.

| System | Training Data | Word error |
|---|---|---|
| Standard | SI | 6.7 |
| Standard | SI+SD | 5.3 |
| +disc | SI | 5.1 |
| +sex | SI+SD | 5.3 |
| +disc | SI+SD | 4.1 |
| +disc+sex | SI+SD | 3.7 |

**Table 4. Summary of Improvements for DECIPHER (Uses February 1989 RM Test Set)**

## Summary

We have shown significant improvements for the DECIPHER speech recognition system by (1) increasing training data size, (2) implementing corrective training, and (3) separating male and female training data. We have combined all three improvements to achieve our best performing system, one that has a word-error rate of 3.7% on DARPA's resource management February 1989 speaker-independent test set.

We believe that the use of a large training set allows significant improvements in speech recognition accuracy, and therefore we advocate using the larger training set as a standard in future system evaluations.

## References


[Bahl88] Bahl, L.R., P.F. Brown, P.V. De Souza, R.L. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," *Proceedings ICASSP-88.*

[Bush87] Bush, Marcia A., and Gary E. Kopec, "Network-Based Connected Speech Recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, October 1987

[Cohen90] Cohen, Michael, Hy Murveit, Jared Bernstein, Patti Price, and Mitch Weintraub, "The DECIPHER Speech Recognition System," *Proceedings ICASSP-90.*

[Hon90] Hon, Hsiao-Wuen, and Kai-Fu Lee, "On Vocabulary-Independent Speech Modeling," *Proceedings ICASSP-90.*

[Jelinek80] Jelinek. F. and R. Mercer, "Interpolated Estimation of Markov Source Parameters from Sparse Data," pp. 381-397 in E.S. Gelsima and L.N. Kanal (editors), *Pattern Recognition in Practice*, North Holland Publishing Company, Amsterdam, the Netherlands.

[Kubala90] Kubala, Francis, Richard Schwartz, and Chris Barry, "Speaker Adaptation from a Speaker Independent Training Corpus," *Proceedings ICASSP-90.*

[Lee89] Lee, K.F., and S. Mahajan, "Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition," Technical Report CMU-CS-89-100, Carnegie Mellon University, January 1989.

[Murveit89] Murveit, Hy , M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein, "SRI's DECIPHER System," *Proceedings of the DARPA Speech and Natural Language Workshop*, February, 1989.

[Pallet89] Pallet, D., Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proceedings ICASSP-89.*

[Paul89] Paul, Douglas, "The Lincoln Continuous Speech Recognition System: Recent Developments and Results," *Proceedings of the DARPA Speech and Natural Language Workshop*, February, 1989.

[Price88] Price, P., W.M. Fisher, J. Bernstein, and D.S. Pallet, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proceedings ICASSP-88.*