# THE LINCOLN CONTINUOUS SPEECH RECOGNITION SYSTEM: RECENT DEVELOPMENTS AND RESULTS[1]

Douglas B. Paul
Lincoln Laboratory, MIT
Lexington, MA 02173

## ABSTRACT

The Lincoln stress-resistant HMM CSR has been extended to large vocabulary continuous speech for both speaker-dependent (SD) and speaker-independent (SI) tasks. Performance on the DARPA Resource Management task (991 word vocabulary, perplexity 60 word-pair grammar) [1] is 3.4% word error rate for SD training of word-context-dependent triphone models and 12.6% word error rate for SI training of (word-context-free) tied mixture triphone models.

## INTRODUCTION

Our earlier development efforts [2,3,4,5,6,7,8,9] centered on improving the SD speaker-stress robustness for both IWR and CSR tasks. Since our IWR database included a normal speech test section, we were able to determine that our enhancements for robustness also improved performance for normally spoken speech (0 errors/1680 test tokens, 105 word vocabulary, multi-style training). An independent test on the TI-20 word database [10] confirmed this normal speech performance with 3 errors out of 5120 test tokens on our first run on this database and no errors after a small amount of development [3]. Our robust CSR database was not useful for determining the large vocabulary normal speech performance.

In order to work on a large vocabulary normal speech CSR task, we switched to the DARPA Resource Management database [1]. The SD portion of this database has 12 speakers with 600 training sentences and 100 development test sentences per speaker. This provided a total of 1,200 test sentences containing 10,242 words. For SI work we used the same development test sentences, but trained on 2,880 sentences from 72 speakers from the SI training portion of the database. (There was an overlap of 8 speakers between the SI and SD training sets making the total of 80 speakers reported in [1].) When additional SI training data was needed, we added the designated "SI development test" data, again avoiding test speaker overlaps, to the designated SI training data. This provided a total 3,990 training sentences from 109 speakers.

The vocabulary of the Resource Management database is 991 words. There is also an "official" word-pair recognition grammar [11]. This grammar is just a list of allowable word pairs without probabilities for the purpose of reducing the recognition perplexity to about 60. (Including the probabilities slightly more than halves the perplexity.)

Working with a single development test set carries a risk of tuning one's system to idiosyncrasies in the test set due to the multiple tests and decisions performed during algorithm development. Methodologies which focus on correcting the individual test set errors are particularly subject to this problem and become, in effect, corrective training [12] on the test set. In contrast, since different test sets have different inherent difficulties, comparisons of two systems using different test sets have a significantly reduced resolution. Therefore, the DARPA program has had several "official evaluation tests", the most recent of which was held in June 88. This test used 25 sentences from each of the 12 speakers (2,546 total words), all of which were new data that had not been used in the development process. These evaluation tests provide the best comparison between systems developed at different sites. The development test data, while less useful for comparing systems developed at different sites, is useful for judging progress over time at a single site, subject to the risk that a

[1] This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government.

later system may enjoy an advantage over an earlier system due to the training to the test set. The results provided below will be identified according to which test set was used: June 88 or development test.

Error rates for these systems will be quoted as "% word error rate" in the text. This number is:

$$\frac{100 * (substitutions + insertions + deletions)}{correct\ number\ of\ words}$$

The detailed results will be found in Table 1 and Table 2. The standard deviations are calculated assuming a binomial distribution for the word error rates.

Table 1: The June 88 System June 88 Test Set Results: % Error Rates.

| Training | Substitution | Insertion | Deletion | Word | Std Dev |
|---|---|---|---|---|---|
| SD | 3.5 | .7 | 1.3 | 5.46 | .45 |
| SI | 8.4 | 2.2 | 2.9 | 13.55 | .68 |
| SI, additional training data | 6.2 | 1.5 | 2.5 | 10.09 | .60 |

Table 2: Development Test Set Results: % Error Rates.

| System | Substitution | Insertion | Deletion | Word | Std Dev | No. Gaussians |
|---|---|---|---|---|---|---|
| SD | | | | | | |
| June 88 | 3.0 | .9 | 1.3 | 5.19 | .22 | – |
| Word Context | 2.1 | .7 | .6 | 3.39 | .18 | – |
| SI | | | | | | |
| June 88 | 7.9 | 2.7 | 2.7 | 13.25 | .34 | 29161 |
| Word Context | 10.2 | 2.9 | 2.5 | 15.56 | .36 | – |
| Tied Mix 10 | 12.3 | 3.9 | 2.7 | 18.87 | .39 | 1471 |
| Tied Mix 20 | 8.1 | 1.9 | 3.2 | 13.19 | .33 | 2941 |
| Tied Mix 40 | 7.8 | 2.2 | 2.6 | 12.62 | .33 | 5881 |
| SI, Additional Training Data | | | | | | |
| June 88 | 6.3 | 1.7 | 2.8 | 10.85 | .31 | – |
| Word Context | 7.9 | 1.8 | 2.2 | 11.86 | .32 | – |

## THE "JUNE 88" CSR SYSTEM

The "June 88" CSR system (which was used for the June 88 DARPA tests) uses a continuous observation HMM with triphone (left and right context-sensitive phone) models [13]. The observation probability density functions are diagonal covariance Gaussians with either a grand (shared) variance or a fixed perceptually-motivated variance. (Both give similar performance on normal speech; however, the perceptually motivated variance appears to be more robust to stress. The grand variance is used on the systems reported here.) The observation vector is a centisecond mel-cepstrum augmented with temporal back differences. The phone models have three states with no state skip transitions. Only one Gaussian per state is used in SD mode. The SI system is identical except that fourth order Gaussian mixtures are used.

The system is trained by an unsupervised bootstrapping procedure. The training data is not time-marked, only its orthographic transcription and a dictionary is required. The initial iterations of the Baum-Welch algorithm are performed using monophone (context-free phone) models from a uniform initial state. This, in effect, automatically marks the data. The monophone models are then used to provide initial values for the (single Gaussian) triphone models and a few more iterations are performed. If mixtures are to be used, minor random perturbations of the single Gaussian mean vectors are used to initialize the Gaussian mixtures and a few final iterations are performed.

During recognition, the system extrapolates (guesses based upon a linear combination of the available triphones) the triphones which were not observed during training. The recognition environment is modeled by adaptive background states. In order to control the relative number of word insertions and deletions, the likelihood is multiplied by a penalty for each word. A Viterbi beam search using a finite state grammar with optional interword silences produces the recognized output.

## RESULTS OF THE JUNE 88 TEST SYSTEM

The SD June 88 test system used word-context-free triphones (i.e., the triphone contexts included word boundaries, but excluded the phone on the other side of the boundary). Since the pronunciation of function words is often idiosyncratic, the triphones for the function words were also word-dependent [14]. The resulting system has 2,434 triphones, 2,413 of which were observed at least once in training and 21 of which were extrapolated by the recognizer. (The same training scripts were used for all 12 SD speakers.) The SD word error rates were 5.46% for the June 88 test set and 5.19% for the development test set.

The SI June 88 test system used the same set of triphones as the SD system. Due to the more varied training data, 2,430 triphones were observed, many from only a few speakers. The word error rate for this system is 13.55% for the June 88 test set and 13.25% for the development test set. The large training set (2,431 observed triphones) produced 10.09% word error rate for the June 88 test set and 10.85% for the development test set.

## WORD BOUNDARY MODELS

The SD system was improved significantly by the addition of word boundary triphone models. (The word-boundary triphones are distinct from word-internal triphones.) In this system, the training data is used twice per Baum-Welch iteration, once to train word-context-free (WCF) models and once to train word-context-dependent models. The same word-internal triphones are used both times. This provides the recognizer with a set of models for the observed word boundaries and a set of WCF models to be used for word boundaries allowed by the grammar but not observed in the training data. This reduces the number of phones extrapolated in the recognizer.

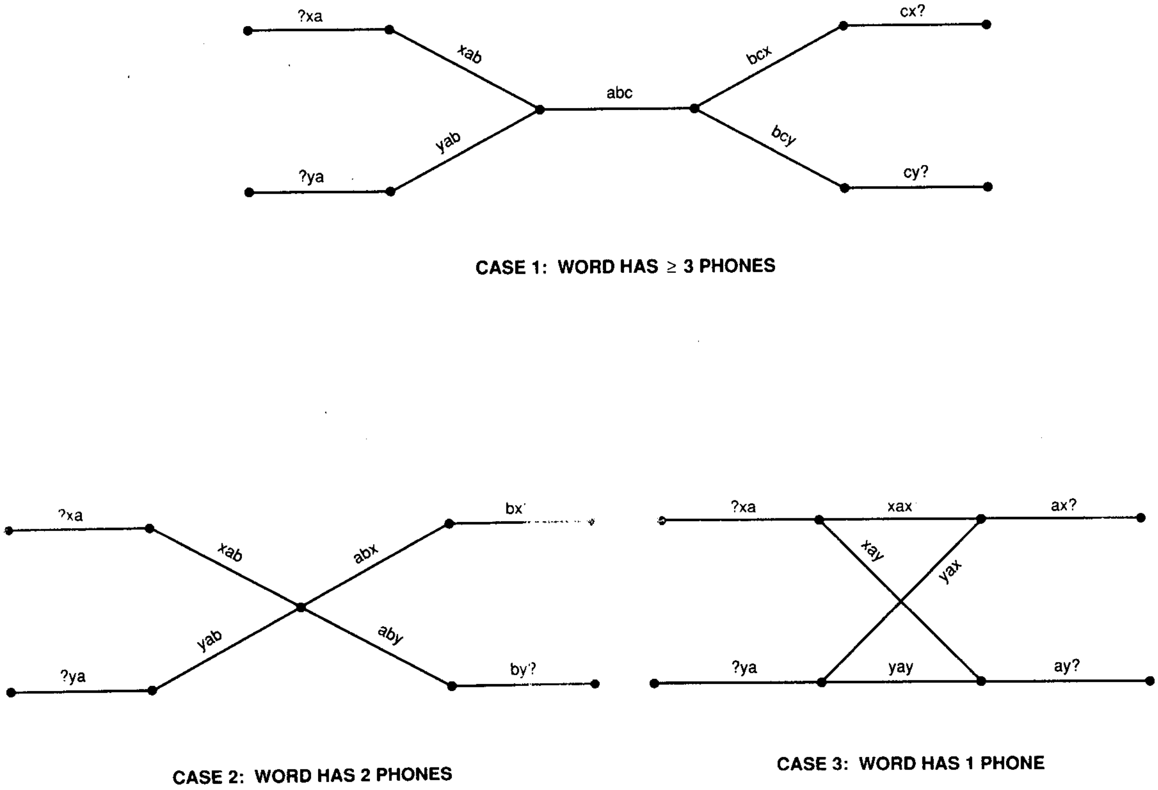
The number of triphones is more than doubled by the added word boundary models. The SD trainer produces 5,993 triphones: 2,413 WCF and 3,580 word context triphones. In addition, the recognizer extrapolates 443 more triphones.

Inclusion of word contexts significantly increases the recognition network complexity. Depending on the number of phones in the word, there are three word topologies which must be covered (Figure 1):

1. Three or more phones: each word end has a fan of initial (final) phones.
2. Two phones: each word end has a list of initial and final phones with a crossbar of interconnections between them.
3. One phone: a crossbar between beginnings and endings with a triphone on each link.

Links between two adjacent words are formed according to the following priority list:

1. Both boundary triphones exist: link them.



CASE 1: WORD HAS ≥ 3 PHONES

CASE 2: WORD HAS 2 PHONES

CASE 3: WORD HAS 1 PHONE

Figure 1: Word-context-dependent word model topologies.

2. Only one of the boundary triphones exist: link to a WCF triphone on the other word.
3. Neither boundary triphone exists: link WCF boundary triphones from both words.

Thus, as more word boundaries are observed in the training data, the system gradually builds from the original WCF system toward a system with full word context models.

The SD development test results for this system showed a significant improvement over the WCF system: 3.39% versus 5.19% word error rate. An earlier system which extrapolated all missing boundary triphones rather the defaulting to WCF triphones did not show an improvement. Thus it is better to use observed WCF triphones rather than extrapolate boundary triphones. The SI results were worse than the WCF system, both with and without the additional training data. The word-context-dependent system appears to be too detailed a model for the available SI training data.

## VARIABLE MIXTURES

Variable order mixtures show a small improvement for the SI task. The number of mixtures for the states in a triphone was chosen by:

$$min(n, sqrt\lfloor(number\ of\ instances\ of\ triphone\ in\ data)\rfloor)$$

This attempts to match the complexity of the distribution to the amount of available training data. It has been tested for $n = 4$ and $n = 8$ with both the normal and augmented training sets. The results are in Table 3. The variable mixtures show an improvement for $n = 4$ for the standard training and $n = 8$ for the augmented training. These results show that the basic idea improves performance but the function is not optimum for choosing the mixture order. If the function were optimum, the best results would be obtained for any large $n$.

Table 3: Variable Mixture Order, Development Test Set Results: % Error Rates.

| System | Substitution | Insertion | Deletion | Word | Std Dev | avg mixture order |
|---|---|---|---|---|---|---|
| SI | | | | | | |
| fixed order 2 | 10.0 | 2.6 | 3.3 | 15.91 | .36 | 2 |
| June 88 (fixed order) | 7.9 | 2.7 | 2.7 | 13.25 | .34 | 4 |
| n=4 | 7.8 | 2.5 | 2.6 | 12.86 | .33 | 3.3 |
| n=8 | 7.9 | 3.3 | 2.3 | 13.47 | .34 | 4.4 |
| SI, Additional Training Data | | | | | | |
| June 88 (fixed order) | 6.3 | 1.7 | 2.8 | 10.85 | .31 | 4 |
| n=4 | 6.1 | 1.5 | 2.8 | 10.37 | .30 | 3.5 |
| n=8 | 6.0 | 2.0 | 2.1 | 10.09 | .30 | 4.9 |

## TIED MIXTURES

A version of tied mixtures [15,16] has been tested and shown to provide a small improvement for the SI task. In this system, each monophone group is given a set of Gaussians. All triphones of each monophone group use mixtures chosen from the same set of Gaussians. The mixture weights for each triphone are independent of all other triphones. This reduces the total number of Gaussians by a significant factor.

Training is again performed using a bootstrapping procedure. After the monophones are trained, small random perturbations of their mean vectors are used to initialize the mixture Gaussians for the monophone group. The triphone weights, along with the parameters of the Gaussians, are then trained with a number of iterations of the Baum-Welch algorithm.

The recognizer used here is the simpler WCF system. Three SI systems were tried using 10, 20, and 40 Gaussians per monophone group. Only the 40 system showed an improvement over the original SI system: 12.62% versus 13.25% development test word error rate. This system also reduced the number of Gaussians by a factor of five. Tied mixtures have not been tried on the SD task.

## SPEAKER GROUPING

Another approach to improving the SI (WCF) performance was tried. The training speakers were segregated by sex and two separate sets of models were trained. The recognizer kept the sets of models separate by using two separate networks. Thus, the system co-recognizes both the speech and the sex of the speaker. Systems which lump both sexes together in training do not discriminate against cross-group spectral matches of individual sounds. Mixtures were not used to save CPU time. The results shown in Table 4 show a significant increase in the error rate.

Table 4: Training Speakers Sex Segregated, Development Test Set Results: % Error Rates.

| System | Substitution | Insertion | Deletion | Word | Std Dev |
|---|---|---|---|---|---|
| SI, no mixtures | | | | | |
| June 88 | 14.0 | 4.3 | 3.7 | 22.02 | .41 |
| Segregated | 17.8 | 3.0 | 7.2 | 28.04 | .44 |

## DISCUSSION AND CONCLUSIONS

Word context modeling reduced the word error rate of the SD system by 35%. This significantly increased both the number of triphones and the complexity of the recognizer. Fully, 39% of the triphones occurred only once in the training data compared to 19% for the WCF system. However, since the system is speaker-dependent and (almost) only Gaussian means were being trained, the system was able to improve in spite of the limited training data.

Word-context modeling did not help the SI system, probably due to insufficient training data. Thirty-two percent of the triphones (mostly word boundary), in contrast to 3.7% for the WCF system, occurred only one or two times in the training data. This is not sufficient to train an SI system. Since many of the word-context-dependent triphones were adapted to only one or two speakers, more damage than good was done. The larger SI training data set reduced the number of single and double occurrence triphones to 22%, which helped, but was not enough to overcome the problem.

Variable order mixtures (WCF) improved the SI results by matching the complexity of the distributions to the amount of training data. This approach required essentially no increase in the complexity of the trainer.

The (WCF) tied mixture system achieved a small improvement over the June 88 SI system. This was probably due to the high-order mixtures of the shared Gaussians. This allowed more detailed modeling where there was sufficient training data while allowing the system to automatically reduce its degrees of freedom where there was insufficient training data by placing very low or zero weights on unneeded mixture components. Training, however, requires so much computation that it will hamper exploration of this class of system.

Grouping the speakers (WCF) yielded sufficiently poor results that mixtures were not tested. The recognizer did appear to correctly identify the sex of the speaker. The reduction in performance may be due to an effect similar to multi-style training [5] which may be enhanced by mixing the sexes during training.

## References

[1] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," ICASSP 88, New York, April 6-9, 1988.

[2] D. B. Paul, R. P. Lippmann, Y. Chen, and C. J. Weinstein, "Robust HMM-Based Techniques for Recognition of Speech Produced Under Stress and in Noise," Speech Tech 86, New York, April 28-30, 1986.

[3] D. B. Paul, "A Speaker-Stress Resistant Isolated Word Recognizer", ICASSP 87, Dallas, April 6-9, 1987.

[4] E. A. Martin, R. P. Lippmann, and D. B. Paul, "Two-Stage Discriminant Analysis for Improved Isolated-Word Recognition," ICASSP 87, Dallas, April 6-9, 1987.

[5] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-Style Training for Robust Isolated-Word Speech Recognition," ICASSP 87, Dallas, April 6-9, 1987.

[6] Y. Chen, "Cepstral Domain Stress Compensation for Robust Speech Recognition," ICASSP 87, Dallas, April 6-9, 1987.

[7] D. B. Paul, "Robust Speech Recognition for Stressful Airborne Environments," Military Speech Tech 87, Washington D. C., November 1987.

[8] D. B. Paul and E. A. Martin, "Speaker Stress-Resistant Continuous Speech Recognition," ICASSP 88, New York, April 11-14, 1988.

[9] E. A. Martin, R. P. Lippmann, and D. B. Paul, "Dynamic Adaptation of Hidden Markov Models for Robust Isolated-Word Speech Recognition," ICASSP 88, New York, April 11-14, 1988.

[10] G. R. Doddington and T. B. Schalk, "Speech Recognition: Turning Theory into Practice," IEEE Spectrum, September 1981.

[11] F. Kubala, Y. Chow, A. Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandergrift, "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database," ICASSP 88, New York, April 11-14, 1988.

[12] L. R. Bahl, P. F. Brown, P. B. de Souza, and R. L. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," ICASSP 88, New York, April 11-14, 1988.

[13] R. M. Schwartz, Y. L. Chow, O. A. Kimball, S. Roucos, M. Krasner, and J. Makhoul, "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," ICASSP 85, Tampa, April 1985.

[14] K. F. Lee and H. W. Hon, "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," ICASSP 88, New York, April 11-14, 1988.

[15] X. D. Huang and M. A. Jack, "Semi-Continuous Hidden Markov Models with Maximum Likelihood Vector Quantization," 1988 IEEE Workshop on Speech Recognition, Arden House, Harriman, NY, May 31-June 3, 1988.

[16] D. Nahamoo, personal communication, June 1988.