The Lincoln Tied-Mixture HMM Continuous Speech Recognizer*

Douglas B. Paul

Lincoln Laboratory, MIT
Lexington, Ma. 02173

Abstract

The Lincoln robust HMM recognizer has been converted from a single Gaussian or Gaussian mixture pdf per state to tied mixtures in which a single set of Gaussians is shared between all states. There were some initial difficulties caused by the use of mixture pruning [12] but these were cured by using observation pruning. Fixed weight smoothing of the mixture weights allowed the use of word-boundary-context-dependent triphone models for both speaker-dependent (SD) and speaker-independent (SI) recognition. A second-differential observation stream further improved SI performance but not SD performance. The overall recognition performance for both SI and SD training is equivalent to the best reported according to the October 89 Resource Management test set. A new form of phonetic context model, the *semiphone*, is also introduced. This new model significantly reduces the number of states required to model a vocabulary.

Introduction

Tied mixture (TM) HMM systems [3, 6] use a Gaussian mixture pdf per state in which a single set of Gaussians is shared among all states:

$$p_i(o) = \sum_j c_{i,j} G_j(o) \qquad (1)$$

$$c_{i,j} \geq 0, \quad \sum_j c_{i,j} = 1$$

where i is the arc or state, G_j is the j^{th} Gaussian, and o is the observation vector. This form of continuous observation pdf shares the generality of discrete observation pdfs (histograms) with the absence of quantization error found in continuous density pdfs. Unlike the non-TM continuous pdfs, TM pdfs are easily smoothed with other pdfs by combining the mixture weights. Unlike discrete observation HMM systems, the Gaussians (analogous to the vector quantizer codebook of a discrete observation system) can be optimized simultaneously with the mixture weights. The training algorithms are identical to the algorithms for training a Gaussian mixture system except the Gaussians are tied across all arcs.

*This work was sponsored by the Defense Advanced Research Projects Agency.

Mixture and Observation Pruning

Computing the full sum of equation 1 is expensive during training and prohibitively expensive during recognition since it must be computed for each active state at each time step. (Because the word sequence is unknown, recognition has many more active states than does training.) Ideally, one would only compute the terms which dominate the sum. However, it requires more computation to find these terms than it does to simply sum them. Two faster approximate methods for reducing the computation exist: mixture and observation pruning.

Mixture pruning simply drops terms that fall below a threshold during training. The weights may then be stored as a sparse array which also saves space. The computational savings are limited during the early iterations of training since only a few terms have been dropped. The final SD distributions are quite sharp (i.e. have only a few terms), but the final SI distributions are quite broad (i.e. have many terms). Thus the savings are limited for SI systems. When the distributions are smoothed with less specific models, they become quite broad again. These difficulties are just computational—there is an even greater difficulty. During training, the parameters of the Gaussians are also optimized which causes them to "move" in the observation space. With mixture pruning, a "lost" Gaussian cannot be recovered. (This was the fundamental difficulty with the earlier version of the system reported in Reference [12].)

Instead of reducing the mixture order, *observation pruning* reduces the computation by computing the sums for all Gaussians whose output probability is above a threshold times the probability of the most probable Gaussian. (Some other sites have used the "top-N" Gaussians [3, 7]. In our system, it gives inferior recognition performance compared to the threshold method.) All of the Gaussians must now be computed, but this is a significant proportion of the computation only in training. (Some pruning is possible. Our exploration of tree-structured search methods showed them to be ineffective because the number of Gaussians is too small

and the observation order is too large.) The amount of computation is now dependent upon the separations of the Gaussian means relative to their covariances and the statistics of the observations. The computational savings were very significant except for the SI second-differential observation stream (discussed later).

Observation pruning does not save space for several reasons. The observation pruned TM systems suffer from the same "missing observation" problem as do the discrete observation systems and therefore no mixture weight can be allowed to become zero. Similarly, recruitment of "new" Gaussians (due to their movement) during training also requires that no mixture weight be allowed to become zero. Both can be accomplished by using full size weight arrays and lower bounding all entries by a small value. Smoothing now causes no organizational difficulty or increase in computation since all mixture weight arrays are full order.

The TM CSR Development

The following development tests were performed using the entire (12 speakers x 100 sentences, 10242 words) SD development-test portion of the Resource Management-1 (RM1) database. Three training conditions were used: speaker-dependent with 600 sentences per speaker (SD), speaker-independent with 2880 sentences from 72 speakers (SI-72) and speaker-independent with 3990 sentences from 109 speakers (SI-109). All tests were performed with the perplexity 60 word-pair grammar (WPG). The word error rate was used to evaluate the systems:

$$\frac{substitutions + insertions + deletions}{correct\ nr\ of\ words}. \quad (2)$$

Line 1 of Table 1 gives the best results obtained from the non-TM Gaussian (SD) and Gaussian mixture (SI) systems [10]. The SD system used word-boundary-context-dependent (WBCD or cross-word) triphone models and the SI systems used word-boundary-context-free (WBCF) triphone models.

The TM HMM systems were trained by a modification of the unsupervised bootstrapping procedure used in the non-TM systems:

1. Train an initial set of Gaussians using a binary-splitting EM to form a Gaussian mixture model for all of the speech data.
2. Train monophone models from a flat start (all mixture weights equal).
3. Initialize the triphone models with the corresponding monophone models.
4. Train the triphone models.

All of the systems described here use centisecond mel-cepstral first observation and time-differential mel-cepstral second observation streams. The Gaussians use a tied (grand) variance (vector) per stream. Each observation stream is assumed to be statistically independent of the other streams. Each phone model is a three state linear HMM. The triphone dictionary also included word dependent phones for some common function words. All stages of training use the Baum-Welch reestimation algorithm to optimize all parameters (the transition probabilities, mixture weights, Gaussian means, and tied variances) simultaneously. The lower bound on the mixture weights was chosen empirically.

The initial observation pruned TM system was derived from the mixture pruned systems described in [12] and gave the performance shown in line 2 of Table 1. It used WBCF triphone models because there was insufficient training data to adequately train WBCD models. Fixed-weight smoothing [15] and deleted interpolation [2] of the mixture weights were tested and the fixed-weight smoothing was found to be equal to or better than the deleted interpolation. (Bugs have been found in both implementations and the smoothing algorithms will require more investigation.) The fixed smoothing weights were computed as a function of the state (left, center, or right), the context (triphone, left-diphone, right-diphone, or monophone) and the number of instances of each phone. The TM system with smoothed WBCF triphone models showed a performance improvement for both the SD and SI trained systems. An additional improvement for both SD and SI systems was obtained by adding WBCD models (table 1, line 3). Until the smoothing was added, we had been able to obtain only slight improvements in the SD systems and no improvement in the SI systems by adding the WBCD models.

Finally, a third observation stream was tested. This stream is a second-differential mel-cepstrum obtained by fitting a parabola to the data within ± 30 msec. of the current frame. It produced no improvement for the SD system, but improved all of the SI systems (table 1, line 4). However, there was a significant computational cost to this stream. Unlike the other observation streams, the number of Gaussians which pass the observation pruning threshold is quite large which slowed the system significantly due to the cost of computing the mixture sums. Increasing the number of iterations of the EM Gaussian initialization algorithm reduced the number of active Gaussians and simultaneously improved results slightly. The computational cost of this stream is still quite large and methods to reduce the cost without damaging performance are still under investigation.

The best systems (starred in table 1) were also tested on the Resource Management-2 (RM2) database. (This database is similar to the SD portion of RM1, except that it contains only four speakers. However, there are 2400 training sentences available for each speaker. The two training conditions are SD-600 (600 sentences) and SD-2400 (2400 sentences). The development tests used 120 sentences per speaker for a total of 4114 words. The RM2 tests (Table 2) showed the SD systems to perform better when trained on more data. One of the speakers (bjw) and possibly a second (lpn) obtained performance which, in this author's opinion, is adequate for opera-

tional use. This is the first time we have observed this level of performance on an RM task. There is still, however, wide performance variation across speakers.

Semiphones

The above best systems all use WBCD triphones. A scan of the 20,000 word Merriam-Webster pocket dictionary yields the following numbers of phones:

	Word Internal	Word Beginning	Word Ending	Cross Word
monophones	43	41	38	-
diphones	1268	602	645	1558
triphones	10788	-	-	49321

(All stress and syllable markings were removed and all possible word combinations were allowed for the cross-word numbers.) This suggests that a large vocabulary system using WBCD triphone models will require on the order of 60K phone models. (Even if the triphones are clustered to reduce the final number [8, 13], all triphones must trained before the clustering process.) These numbers assume no function word or stress dependencies. (A variety of other context factors have also been found to affect the acoustic realization of phones [4].) While this number is not impossible—the Lincoln SI-109 WBCD system has about 10K triphones and CMU used up to 38K triphones in their vocabulary independent training experiments [5]—it is rather unwieldy and would require large amounts of data to train the models effectively. (60K triphones would require about 280M mixture weights and accumulation variables in the Lincoln SI system.)

One possible method of reducing the number of models is the *semiphone*, a class of phone model which includes classic diphones and triphones as special cases. (A classic diphone extends from the center of one phone to the center of the next phone. In a triphone based system, a diphone is a left or right phone-context sensitive phone model.) The center phone of a three section semiphone model of a word with the phonetic transcription /abc/ would be:

$$a_r\text{-}b_l\text{-}b_m \quad b_l\text{-}b_m\text{-}b_r \quad b_m\text{-}b_r\text{-}c_l$$

where l denotes the left part, m the middle part, and r the right part. As shown here, each section is written as a left and right context dependent section (i.e. a "tri-section"). Thus the middle part always has the same contexts and is therefore only monophone dependent. The left (and right) sections are dependent upon the middle part, which is always the same, and a section of the adjacent phone. Thus the left part is similar to the second half of a classic diphone, the center part is monophone dependent, and the right part is similar to the first half of a classic diphone. (In fact, we implemented the scheme using the current triphone based systems simply by manipulating the dictionary.) If the middle part is dropped, this scheme implements a classic diphone system and if the left and right parts are eliminated it reverts to the standard triphone scheme.

One of the advantages of this scheme is a great reduction in the number of models. For the above dictionary, the three section model has 5695 phones. (This number was derived from the above table and is therefore not quite correct since the single phone words were not treated properly. However, the number is sufficiently accurate to support the following conclusions.) If the semiphone system has one state per phone and the triphone system has three states per phone, each word model will have the same number of states (for a given left and right word context), but the semiphone system will have 5695 unique states to train and the triphone system will have 180K unique states to train.

Semiphones avoid one of the difficult aspects of cross-word triphones—the single phone word. A single phone word requires a full crossbar of triphones in the recognition network [11]. The semiphone approach splits the single phone into a sequence of two or more semiphones and simply joins the apexes of a left fan and a right fan for a two semiphone model or places the middle semiphone between the fans for a three semiphone model [11].

A final advantage of the semiphone approach over the classic diphone approach is the organization. The units are organized by the phone. This is a more convenient organization for smoothing and also makes the word endpoints explicitly available for word endpointing or any word based organization of the recognizer.

Our current implementation of this scheme has not yet addressed smoothing the mixture weights of the semiphones, so the results-to-date can only compare unsmoothed semiphone systems with smoothed triphone systems. Line 1 of Table 3 repeats the corresponding entries for two smoothed triphone systems from Table 1 for comparison with the semiphone systems. Line 2 is an unsmoothed three-section semiphone system with one state per semiphone. For both training conditions, the number of unique states was reduced by about a factor of five. The difference in performance between the systems is commensurate with the difference between smoothed and unsmoothed triphone systems. Line 3 is equivalent to a classic diphone system with two states per semiphone and thus four states per phone rather than three states per phone as in the preceding systems. This system has twice as many states as the other semiphone system and yields equivalent performance. While the semiphone systems do not currently outperform the triphone systems, they bear further investigation.

The October 89 Evaluation Test Set

At the time of the October 89 meeting, the mixture pruned systems were not showing improved performance over the best non-TM systems and therefore non-TM

systems were used in the evaluation tests. The best observation pruned systems (starred in Table 1) were tested using the October 89 test set in order to compare them to the results obtained at the other DARPA sites. The results are shown in Table 4. These results are not statistically distinguishable from best results reported by any site at the October 89 meeting [14].

The June 90 Evaluation Tests

The best TM triphone systems (starred in Table 1) were used to perform the evaluation tests. Both systems used WBCD triphones with fixed weight smoothing. The SD systems used two observation streams and the SI-109 system used three observation streams. The results are shown in Table 5.

Conclusion

The change from mixture pruning to observation pruning has eliminated the Gaussian recruitment problem. The change increased the data space requirements, but provided a better environment for mixture weight smoothing and reduced the computational requirements for both training and recognition. Including fixed-smoothing-weight mixture-weight smoothing improved performance on both SD and SI trained systems and allowed the use of WBCD (cross-word) triphone models.

Testing on the RM2 database showed that our systems developed on the RM1 database transferred without difficulty to another database of the same form. It also showed that our SD systems will provide better performance when given more training data (2400 sentences) than is available in the RM1 database (600 sentences). Operational performance levels were obtained on one or two of the (four) speakers.

We found a simpler context-sensitive model—the semiphone—to produce similar recognition performance to the (by now) traditional triphone systems. These models, which include the classical diphone as a special case, significantly reduce the number of states (or observation pdfs) which must be trained. The semiphone model will require further development and verification but it may be one way of simplifying our systems. Since the number of semiphones required to cover a 20,000 word dictionary is significantly less than the number of triphones required to cover the same dictionary, they may be a more practical route to vocabulary independent phone modeling than one based upon triphones.

References

[1] S. Austin, C. Barry, Y. L. Chow, A. Derr, O. Kimball, F. Kubala, J. Makhoul, P. Placeway, W. Russell, R. Schwartz, and G. Yu, "Improved HMM Models for High Performance Speech Recognition," Proceedings DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October, 1989.

[2] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-5, March 1983.

[3] J.R. Bellagarda and D.H. Nahamoo, "Tied Mixture Continuous Parameter Models for Large Vocabulary Isolated Speech Recognition," Proc. ICASSP 89, Glasgow, May 1989.

[4] F. R. Chen, "Identification of Contextual Factor For Pronunciation Networks," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.

[5] H. W. Hon and K. F. Lee, "On Vocabulary-Independent Speech Modeling," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.

[6] X. D. Huang and M.A. Jack, "Semi-continuous Hidden Markov Models for Speech Recognition," Computer Speech and Language, Vol. 3, 1989.

[7] X. Huang, K. F. Lee, and H. W. Hon, "On Semi-Continuous Hidden Markov Modeling," Proc. ICASSP90, Albuquerque, New Mexico, April 1990.

[8] K. F. Lee, Automatic Speech Recognition: The Development of the SPHINX System, Kluwer Academic Publishers, 1989.

[9] K. F. Lee, Presentation at DARPA Speech and Natural Language Workshop, October 1989.

[10] D. B. Paul, "The Lincoln Continuous Speech Recognition System: Recent Developments and Results," Proceedings DARPA Speech and Natural Language Workshop, February 1989, Morgan Kaufmann Publishers, February 1989.

[11] D. B. Paul, "The Lincoln Robust Continuous Speech Recognizer," Proc. ICASSP 89., Glasgow, Scotland, May 1989.

[12] D. B. Paul, "Tied Mixtures in the Lincoln Robust CSR," Proceedings DARPA Speech and Natural Language Workshop, Morgan Kaufmann Publishers, October, 1989.

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

[14] Proceedings DARPA Speech and Natural Language Workshop, October 1989, Morgan Kaufmann Publishers, October, 1989.

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

Table 1. RM1 Development Test Results using triphone models. The standard deviations are computed for the best result in each column.

% Word Error Rates with WPG

				SD		SI-72		SI-109	
	System	Gaussians	Smoothing	WBCF	WBCD	WBCF	WBCD	WBCF	WBCD
1.	Non-TM	many	n	5.2	3.0	12.9	-	10.1	-
2.	TM-2	2x257	n	4.3	-	14.0	-	-	-
3.	TM-2	2x257	y	3.3	1.7*	11.3	9.0	10.4	7.8
4.	TM-3	3x257	y	-	1.8	9.5	7.2	8.5	5.6*
Binomial standard deviations				(.18)	(.13)	(.29)	(.26)	(.27)	(.23)

*=evaluation test (best) systems

Table 2. RM2 Development test results using the best (starred) systems of Table 1.

% Error Rates with WPG

	SD-2400		SD-600	
Speaker	word (sd)	sentence	word (sd)	sentence
bjw	.2	1.7	1.4	10.0
jls	.7	5.0	3.8	21.7
jrm	2.8	16.7	4.3	26.7
lpn	.4	3.3	2.5	15.0
avg	1.0 (.16)	6.7	3.0 (.27)	18.3

System: TM2, Gaussians: 2x257, Smoothed, Triphone models

Table 3. Semiphone development tests using the RM1 database. The standard deviations are computed for the best result in each column. Line 1 is the best (starred) results from Table 1.

% Word Error Rates with WPG

				Sections	States per	SD		SI-109	
	System	Gaussians	Smoothing	per Phone	Section	States	Errors	States	Errors
1.	TM-2,triphone	2x257	y	1	3	17979	1.7	24201	7.8
2.	TM-2,semiphone	2x257	n	3	1	3793	2.2	4372	9.5
3.	TM-2,semiphone	2x257	n	2	2	7286	2.3	-	-
Binomial standard deviations							(.13)		(.27)

Table 4. Results for the best (starred) systems of Table 1 using the October 89 evaluation test (RM1) data.

% Word Error Rates (std dev) with WPG

System	Gaussians	Smoothing	SD	SI-109
TM-2	2x257	y	2.6 (.31)	-
TM-3	3x257	y	-	5.9 (.45)
Best from any site			2.5 [1]	6.0 [9]

Table 5. The June 1990 Evaluation test results using triphone based systems on the RM2 database. The systems are the best (starred) systems of Table 1.

% Word Error Rates (std dev)

		Word-pair Grammar (p=60)					No Grammar (p=991)*				
	Training	sub	ins	del	word (sd)	sent	sub	ins	del	word (sd)	sent
TM-2	SD-2400	.9	.2	.4	1.51 (.19)	11.0	3.3	.8	.9	4.89 (.34)	28.8
TM-2	SD-600	1.7	.5	.9	3.09 (.27)	20.0	8.3	2.2	2.2	12.66 (.52)	58.3
TM-3	SI-109	3.8	.7	1.3	5.86 (.37)	31.9	16.5	2.1	4.4	22.92 (.66)	74.6

* Homonyms equivalent