TIED MIXTURES IN THE LINCOLN ROBUST CSR[1]

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

ABSTRACT

HMM recognizers using either a single Gaussian or a Gaussian mixture per state have been shown to work fairly well for 1000-word vocabulary continuous speech recognition. However, the large number of Gaussians required to cover the entire English language makes these systems unwieldy for large vocabulary tasks. Tied mixtures offer a more compact way of representing the observation pdf's. We have converted our independent mixture systems to tied mixtures and have obtained mixed results: a 13% improvement in speaker-dependent recognition without cross-word triphone models, but no improvement in our speaker-dependent system with cross-word boundary triphone models or in our speaker-independent system. There is also a reduction in CPU requirements during recognition—but this is counter-balanced by an increase during training. This paper also includes a comment on the validity of the DARPA program's evaluation test system comparisons.

INTRODUCTION

Single Gaussian per state speaker-dependent (SD) HMM recognizers and low-order Gaussian mixture per state speaker-independent (SI) HMM recognizers have been shown to work fairly well for 1000-word vocabulary, continuous speech recognition [10,11]. However, a SD system would require about 30,000 Gaussians to cover the word-internal triphones of English and a SI system would require at least 100,000. The strategy of one or more individual Gaussians per state is appropriate for small vocabulary systems, but becomes unwieldy for large vocabulary systems. Interpolation is often required to cluster models, smooth models, or to predict models which are not observed in training—but there is no clean strategy for interpolating independent Gaussian mixtures—either the mean(s) are changed or the mixture order increases each time another model is included into an interpolated model.

Tied mixtures [3,2,4] offer a solution for these problems while retaining a basic continuous observation HMM system. (Gaussian tied mixtures are mixtures which share a common pool of Gaussians.) They are mixtures, and thus avoid the unimodal distribution limitation of single Gaussians. Unlike independent mixtures, they interpolate well by interpolating the weights of the corresponding Gaussians. And since the pool of Gaussians is of a given size, a mixture order cannot exceed this size. In effect, they form a middle ground between the histograms of discrete observation systems and non-tied-mixture systems. Tied mixtures can also be viewed as a discrete observation system modified to allow a simultaneous match to many templates with the degree of template match included. In contrast to the discrete observation system, there is no quantization error and the "templates"(Gaussians) can be jointly optimized with the rest of the HMM.

[1]This work was sponsored by the Defense Advanced Research Projects Agency.

TIED MIXTURES

A tied mixture HMM system simply substitutes a mixture with shared Gaussians for the observation pdf in a continuous observation HMM:

$$b_i(o) = \sum_j w_{ij} N_j(o) \tag{1}$$

$$\sum_j w_{ij} = 1 \tag{2}$$

where i is the state (or arc), b is the observation pdf, o is an observation vector, w is the weight, and N_j is the set of shared Gaussians. The forward-backward re-estimation procedure is identical to the procedure for independent mixtures except the Gaussians are tied. (The equations can be derived trivially from the well known independent mixture case. They are presented in [3,2,4] and are not repeated here.) In general, all Gaussians are used by all states, but in practice most of the weights are set to zero by the training procedure. However, the average mixture order can be very high during the early phases of training.

TIED MIXTURE SYSTEMS AT OTHER SITES

Several other sites have experimented with tied mixture HMM recognizers [3,13,2,4]. However, the initial parameters for training these systems have been derived from existing discrete observation HMM systems. The initial Gaussian means and covariances were derived from the templates of the vector quantizer, and the mixture weights were initialized from the observation probability histograms. All of these sites reported moderate performance improvements over their discrete observation systems. The work reported here does not bootstrap from any discrete observation system. This provides some additional freedom in training which may influence the final recognition performance.

THE TESTS

All system tests reported here were performed on the DARPA Resource Management (RM) database [12]. The SD system was trained on the designated 600 training sentences per speaker, and the SI system was trained on either 72 designated training speakers × 40 sentences per speaker = 2880 sentences (SI-72) or the SI-72 + 37 development test speakers × 30 sentences per speaker = 3990 sentences (SI-109). Only 72 of the 80 SI training and 37 of the 40 SI development test speakers could be used because the other speakers are contained in the test set. Except for the evaluation tests, the test set in all cases was all 100 development test sentences per speaker for the 12 SD speakers. These 1200 sentences contain 10242 words. The word error rate is:

$$\frac{(substitutions + insertions + deletions)}{correct\ nr\ of\ words} \tag{3}$$

The recognition development test results quoted in the text and in Table 1 are percent word error rate with the perplexity 60 word-pair grammar.

THE TIED MIXTURE SYSTEMS AND EXPERIMENTS

The systems reported at the February 1989 DARPA meeting (the "Feb89" systems) [10,11] were a single Gaussian per state HMM with word (boundary) context-dependent (WCD) triphones for SD, and a variable order Gaussian mixture per state HMM with word (boundary) context-free (WCF) triphone models. All Gaussians used a single-tied (grand) diagonal covariance matrix. The observation vector is a 10 ms mel-cepstrum augmented with a temporal difference ("delta") mel-cepstrum. The development test performances are shown in Table 1.

The tied-mixture systems were initialized by a modification of our monophone bootstrapping procedure [10]. As in the Feb89 systems, single Gaussian monophone (context independent phone) models were trained from a "flat" start (all phones identical) and used to initialize single Gaussian triphone models. This produced about 7200 (one per state) Gaussians with a tied (grand) variance. The means of these Gaussians were treated as observations and clustered down to 256 clusters by a binary-splitting k-means algorithm. (The tied variance was used but not altered during clustering.) The mixture weights were initialized by computing the Gaussian probability of the cluster mean given the state and then normalizing according to Eq. 2. All parameters (transition probabilities, distribution weights, Gaussian means, and tied variance) were trained. (Each stage of training used the forward-backward algorithm.) If a mixture weight became less than a threshold, the component was removed from the mixture. Thus the mixtures were automatically pruned in response to the training data to reduce the computation. Average mixture orders were initially very high, but were reduced significantly by the end of training.

The first tied mixture system used only mel-cepstral observations, WCF triphone models, and 256 Gaussians. (Unless otherwise noted, all of the following systems use WCF triphone models.) Results for SD (5.5% word errors) were very similar to the corresponding Feb89 system (5.2%), but the SI-72 performance was significantly degraded: 26.2% word errors vs. 12.9% for the Feb89 system. The reduced performance without the delta mel-cepstral parameters was not unexpected. However, the number of Gaussians was reduced from 24,000 for SI-72 and 7200 for SD to 256.

Delta mel-cepstral parameters were then returned to the system by augmenting the observation vector. The performance on the SD task decreased to 6.1% word errors, but the SI-72 task improved to 17.2% word error rate. Including the delta parameters changed the relation between the mel-cepstral and delta mel-cepstral observations for the SD system. In the single Gaussian case, the diagonal covariance matrix treated the mel-cepstral and the delta mel-cepstral observations as statistically independent. However, the mixture weights induced a relation between the two parameter sets. (They were already related in the SI system due to the independent mixtures.) Increasing the number of Gaussians to 512 to increase the system's ability to model the correlation between the mel-cepstral and delta mel-cepstral parameters improved the SD performance to 5.0% word errors but had no effect on SI-72: 17.1% word errors. It appears that there was insufficient data to train the correlations or still an insufficient number of Gaussians to model the correlations in the SI task.

A number of other sites [14,5,6], for example, have improved performance with limited training data by separating different parameters into separate observation streams and multiplying their respective observation probabilities to force the HMM to treat them as if they were statistically independent. Therefore, the mel-cepstra and the delta mel-cepstra were split into separate observation streams:

$$b_i(o_c, o_d) = b_{c,i}(o_c) * b_{d,i}(o_d) \quad (4)$$

where c denotes mel-cepstrum and d denotes the delta mel-cepstrum. This maintained the performance on the SD task (5.0% word errors) and further improved the performance on the SI-72 task to 14.7% word errors.

Next, the training procedure was modified by, instead of clustering the means of the Gaussians, clustering a subset of the training data, again using a binary-splitting k-means algorithm. It was hoped that this would provide an initialization with better representation of outliers which might have been suppressed by the single Gaussians. This change resulted in improvements in both tasks: the SD error rate went down to 4.7% and the SI-72 error rate went down to 13.7%

A variation in the training procedure of the "kt" systems was tested. It was feared that the high-frequency triphones were dominating the Gaussian means in the early iterations of training causing damage to the modeling of low-frequency triphones. Therefore, the Gaussian means were not trained until the weights had settled. This was intended to protect the Gaussian means until the phone models had become very specific. No improvement was found.

To fully test for outliers, the system was initialized with a set of Gaussians formed by binary-splitting k-means clustering a subset of the training data using the perceptually-motivated weighting [8,9] (which was again not altered during clustering). The system was started with flat start tied mixture monophones (maximum order mixtures with all weights equal). These monophone models were used to bootstrap the triphone models, again using the forward-backward algorithm at each stage. These "ks" systems provided the best performance for the SD task (4.5% word errors), but failed to improve on the SI-72 task (15.3% word errors), probably due to the slight smoothing induced by the old initialization. This SD performance is better than the corresponding Feb89 system with WCF triphone models (5.2%).

None of the above systems used word context (boundary) modeling. The "kt" system was tested on the SD task using word context-dependent models. The performance (4.0% word errors) was better than the WCF system (4.7% word errors), but was not better than the Feb89 SD systems with WCD models (3.0% word errors). The tied mixture system appears to require more training data than does a single Gaussian per state system.

The above systems do not have any smoothing on the mixture weights. A preliminary attempt to use deleted interpolation across phonetic contexts [1,5] caused a slight increase in the error rate of an SI-72 system.

DISCUSSION

The changeover to tied mixtures has achieved better performance than the WCF SD system. The improvement due to adding the delta mel-cepstral observations was quite small (5.5 to 5.0% word error rate) compared the improvement on the SI-72 task (26.7 to 14.7% word error rate). The SD improvement found here is similar to that achieved in a similar test with a single Gaussian per state WCF SD system. In contrast, BBN [14] achieved a dramatic improvement by adding delta observations to their SD system. It is not obvious why the effect of delta observations should be so variable.

The net improvement in the WCF SD system but not the WCD SD system and the SI systems in the context of the changeover to tied mixtures suggests the need for smoothing of the weights. (The WCF systems have about 2400 triphones and the SD WCD has about 6000 triphones.) The mixture weights in the tied mixture systems give more degrees of freedom in spectral matching than single Gaussians or low-order independent mixtures and, therefore, require more training data or smoothing to be effective. The "kt" system was SI-109 trained which, as expected, improved the results (13.7 to 11.2% word error). However, it still did not outperform the independent mixture system (10.1% word errors). The attempt to use deleted interpolation to smooth the weights of an SI-72 system failed for reasons that are as yet unknown. (It might be a defect in the details of our technique or just a bug in our program.) Smoothing will require re-examination.

Tied mixture systems are very compute-intensive. Some of the other tied mixture systems have attempted to reduce computation by limiting the number of "active" Gaussians to a few with the highest probability [3,4,13]. The systems used here dynamically reduced the mixture order by removing components if their weights fell below a threshold. This resulted in long iteration times early in the training when the mixtures were still of a high order, but the later iterations proceeded at a reasonable pace. The recognizer, of course, saw only the lowest order mixtures from the final iteration of training. The net effect is an approximate doubling of the training time over the Feb89 systems and a halving of the recognition times. Since most experiments require both training and recognition, the total experiment time was significantly increased. A changeover to limiting the number of active Gaussians may reduce the training time.

Our work with tied mixtures has shown promise of improved performance. There are still a number of issues to be examined or re-examined, and we will continue working on tied mixture HMM recognition systems.

COMMENT ON COMPARATIVE EVALUATION TESTING

The standard deviations in Table 1 are computed, assuming a binomial distribution. This standard deviation is very optimistic—it assumes that all conditions are equal and that all errors are independent. If it has any validity, it is valid only for comparisons of very similar (preferably minimal pair) systems tested on exactly this data. A standard deviation computed across speakers is much higher—see Table 2 for some comparative values. This standard deviation gives an idea of the confidence one would have in predicting the recognition performance of a new speaker. The high across-speaker standard deviation also suggests that comparisons between systems are highly dependent upon the speaker set used for the comparison tests.

Using the SD February 89 evaluation test data, we compared the best six (of twelve) speaker lists from both the Lincoln and BBN systems, and found only three (50%) speakers common to both lists. This was true with and without the word-pair grammar. A similar comparison on the best five (of ten) speakers yields a list intersection averaged over all site pairs with grammar and all site pairs without grammar of 63% for the SI-72 task (Lincoln, MIT, and SRI systems) and 73% for the SI-109 task (CMU, Lincoln, and SRI systems). In general, since the SI training only uses a moderate number of speakers, a higher correlation between the best lists is expected for SI than for SD because the SI test speakers may be more or less similar to the training speakers; whereas, no such systematic variation exists for the SD tests.

This analysis is ad-hoc, but the high across-speaker standard deviation and the poor agreement in

the best speaker lists suggest a weakness in our current inter-site system comparison procedures. Strengthening our inter-site comparisons cannot be achieved just by more powerful tests for comparing two sets of results for the same speaker—much larger test speaker sets and tests which take into account the inter-speaker variation are required. SRI has made a comparison based upon summing the per-speaker comparisons between two systems and reached a similar conclusion [7]. In practice, we may not be able to collect adequate data from, for example, 100 speakers for SD training and testing, but 100 test speakers for SI testing is quite practical.

EVALUATION TESTS RESULTS

Immediately after the February 89 meeting, a bug was found in the recognition network generation software for the WCD models. (This bug only affected our SD WCD system.) The fix was not a change in concept, only a correction of the implementation. The development test results are shown in Table 1 as "'Feb89 WCD with bug" and "Feb89 WCD". The comparisons in this paper have been made with the "Feb89 WCD" system because it is (barring other bugs) the implementation of the system described in the talk and paper [10]. These tests were rerun and filed with NIST. A summary of the February 89 SD evaluation test results with and without the bug are shown in Table 3.

Since our tied mixture systems have not shown better performance than our (fixed) SD "Feb89 WCD" single Gaussian per-state system and our "Feb 89" independent mixture SI systems, the Feb89 systems are being used for the evaluation tests. A summary of the results is shown in Table 4.

There appears to be a bias in the SD test data relative to other test data—the SD tests show significantly fewer insertions than deletions. The effect is very strong for the with grammar test where 10 of the 12 speakers showed no insertion errors. In contrast, only one speaker showed no deletion errors. The February 89 tests of the same SD system show a balance between the two forms of errors (Table 3). (There is a similar, but weaker, bias in the no grammar SD case. This weaker bias would not be noteworthy if the with grammar case did not call attention to it.) The insertion penalty, which controls this trade-off, was set for minimum word error rate on the development test data. Usually, this minimum occurs when the insertion and deletion error rates are similar. The skew observed here is large enough that the SD word error rate would probably be reduced if the insertion penalty were adjusted to match this test data. (The excess of deletion errors in the current SI tests also occurred in the February 89 tests and is, therefore, not noteworthy.)

References

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech Recognition," PAMI-5, No. 2, March 1983.

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

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

[4] X. D. Huang, H. W. Hon, and K. F. Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition with Semi-Continuous Hidden Markov Models," Eurospeech 89, Paris, September 1989.

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

[6] H. Murveit and M. Weintraub, "1000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models," ICASSP 88, New York, April 1988.

[7] H. Murveit, M. Cohen, P. Price, G. Baldwin, M. Weintraub, and J. Bernstein, "SRI's DECIPHER System," Proceedings DARPA Speech and Natural Language Workshop, Philadelphia, February 1989.

[8] D. B. Paul, "A Speaker-Stress Resistant Isolated Word Recognizer," ICASSP 89, Dallas, Texas, April 1987.

[9] D. B. Paul, "Speaker Stress-Resistant Continuous Speech Recognition," ICASSP 88, New York, April 1988.

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

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

[12] 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 1988.

[13] R. Schwartz, personal communication, September 1989.

[14] R. Schwartz, C. Barry, Y. L. Chow, A. Derr, M. W. Feng, O. Kimball, F. Kubala, J. Makhoul, and J. Vendegrift, "The BBN BYBLOS Continuous Speech Recognition System," Proceedings DARPA Speech and Natural Language Workshop, Philadelphia, February 1989.

TABLE 1
DEVELOPMENT TEST RESULTS WITH p=60 WORD-PAIR GRAMMAR

System	k-means on	% Word Errors (% std dev) Training Condition SD	SI-72	SI-109
Feb89 WCF	–	5.2 (.2)	12.9 (.3)*	10.1 (.3)*
single observation stream				
cep	tri	5.5 (.2)	26.2 (.4)	
cep+delta	tri	6.1 (.2)	17.2 (.4)	
cep+delta, 512 Gaussians	tri	5.0 (.2)	17.1 (.4)	
multiple observation streams				
cep+delta	tri	5.0 (.2)	14.7 (.4)	
cep+delta: "kt"	obs	4.7 (.2)	13.7 (.3)	11.2 (.3)
cep+delta: "ks"	start	4.5 (.2)	15.3 (.4)	
word boundary context-dependent				
Feb89 WCD with bug	–	3.4 (.2)*		
Feb89 WCD	–	3.0 (.2)**		
cep+delta: WCD "kt"	obs	4.0 (.2)		

All tests use the SD Development Test set: 12 SD speakers, 100 sentences per speaker, 10242 total words. The std dev assumes a binomial distribution. The "k-means on" column gives the data used by the k-means algorithm to create the set of Gaussians used by the tied mixtures. All tied mixture systems use 256 Gaussians unless otherwise noted.

* Feb89 official test systems
** fixed Feb89 SD test system
k-means codes (see text for complete description):
start = k-means of data, tied mixtures at all stages of training
tri = single Gaussian bootstrap, k-means of triphone means
obs = single Gaussian bootstrap, k-means of observation data
– = not a tied mixture system

TABLE 2
COMPARATIVE DEVELOPMENT TEST STANDARD DEVIATIONS

System	Word Errors	Standard Deviations	
		Binomial	Across Speaker
SD Feb89 WCD	3.0%	.17%	1.03%
SD Feb89 WCF	5.2%	.22%	1.16%
SI-72 Feb89	12.9%	.33%	4.97%
SI-109 Feb89	10.1%	.30%	4.68%

TABLE 3
SUMMARY OF FEBRUARY 89 EVALUATIONS TEST RESULTS
WITH WCD MODEL RECOGNIZER BUG FIXED

System	% Word Error Rates									
	Word-Pair Grammar (p=60)					No Grammar (p=991)*				
	sub	ins	del	word (sd)	sent	sub	ins	del	word (sd)	sent
Feb89 SD WCD, bug	2.7	1.0	.5	4.2 (.4)	28.0	8.4	2.9	2.0	13.2 (.7)	60.3
Feb89 SD WCD	2.5	.6	.6	3.6 (.4)	25.7	8.6	2.5	2.1	13.1 (.7)	60.0

* Homonyms equivalent
Binomial standard deviations

TABLE 4
SUMMARY OF OCTOBER 89 EVALUATION TEST RESULTS

	% Word Error Rates									
	Word-Pair Grammar (p=60)					No Grammar (p=991)*				
System	sub	ins	del	word (sd)	sent	sub	ins	del	word (sd)	sent
October 89 test set										
Feb89 SD WCD	2.5	.1	1.0	3.6 (.4)	24.0	9.9	1.5	2.6	14.0 (.7)	63.7
Feb89 SI-72	7.2	1.6	3.2	12.0 (.6)	52.7	21.9	4.4	8.2	34.5 (.9)	91.0
Feb89 SI-109	6.1	1.9	2.8	10.8 (.6)	44.7	18.9	3.1	7.8	29.8 (.9)	88.3
"Retest" test set										
Feb89 SD WCD						8.9	2.5	1.7	13.1 (1.0)	60.0
Feb89 SI-109						17.9	3.1	6.0	26.9 (1.2)	78.7

* Homonyms equivalent
Binomial standard deviations