BATCH, INCREMENTAL AND INSTANTANEOUS ADAPTATION TECHNIQUES FOR SPEECH RECOGNITION

G. Zavaliagkos† *R. Schwartz* *J. Makhoul*

BBN Systems and Technologies, Cambridge, MA 02138
†Northeastern University , Boston MA 02115
GZaval@bbn.com

ABSTRACT

We present a framework for Maximum A Posteriori adaptation of large scale HMM speech recognizers. In this framework, we introduce mechanisms that take advantage of correlations present among HMM parameters in order to maximize the number of parameters that can be adapted by a limited number of observations. We are also separately exploring the feasibility of instantaneous adaptation techniques. Instantaneous adaptation attempts to improve recognition on a single sentence, the same sentence that is used to estimate the adaptation.
In a nutshell, we show that sizable gains (20-40% reduction in error rate) can be achieved by either batch or incremental adaptation for large vocabulary recognition of native speakers. The same techniques cut the error rate for recognition of non-native speakers by factors of 2 to 4, bringing their performance much closer to the native speaker performance. We also demonstrate that good improvements in performance (25-30%) are realized when instantaneous adaptation is used for recognition of non-native speakers.

1. INTRODUCTION

State-of-the-art speaker independent (SI) HMM recognizers use tens of hours of speech from many speakers to estimate their parameters. The resulting models may perform well, but they do not represent the true distribution of *any* of the training speakers. Furthermore, these models suffer serious degradations when tested on speakers/environments not represented in the training. A speaker dependent (SD) system would, in principle, outperform a SI system, but it is usually impractical to collect the hours of speech required to train it from a single speaker. Instead, adaptation techniques are called for to move the SI model closer to the model of the speaker in question and improve performance.

In general there are three strategies on how to adapt the HMM: (i) batch adaptation, where we collect a limited amount (minutes) of enrollment data and use it to train the adaptation; (ii) incremental adaptation, where the system adapts unsupervised every sentence the user says and uses the adapted model for the next sentence and (iii) instantaneous adaptation, where we attempt to improve recognition on the same data that is used to estimate the adaptation transformation. Instantaneous adaptation is especially useful in applications where there is a very brief interaction between the speaker and the system; the use of even the smallest amount of enrollment data is then inconceivable, and incremental adaptation is not given enough time to improve significantly.

Our algorithms for supervised and incremental adaptation tasks are based on the Maximum A Posteriori (MAP) estimation of HMM parameters [1, 2]. MAP provides an optimal solution to the problem of combining the prior knowledge of the SI system with the data from the new speaker, and it has the advantage of asymptotically converging to the SD model. However, MAP changes only the parameters for which data are available. To ameliorate this problem we introduce two generalization mechanisms: (i) the application of an Extended MAP (EMAP) estimation that takes into account correlations present across HMM parameters [3], and (ii) smoothing new estimates of detailed models with the estimates of artificially introduced more general models.

Instantaneous adaptation has been rarely treated in the literature so far. The problem is harder than other adaptation tasks because we have very little data to change our model (one sentence) and we are testing on the same data used to estimate the adaptation. Consequently algorithms that require a lot of adaptation data to become effective are destined to fail. Similarly, transformations that have many degrees of freedom (and hence "learn" the adaptation data quickly) will simply repeat the errors that are present in the recognizer output which is used to supervise the adaptation. The approach we follow is to tie all the parameters to phoneme classes and then use EMAP and the Probabilistic Spectral Mapping (PSM) [4] to re-estimate them.

We evaluate our work with the BBN BYBLOS Phonetically Tied Mixture (PTM) HMM, and we present results of the 5K word tests of the ARPA Wall Street Journal (WSJ) Corpus [5]. Particular emphasis is given to the recognition performance of non-native speakers of American English. These speakers, having on the average 4 to 6 *times* the error the native speakers have, present a serious obstacle to the widespread use of commercial speech recognition systems.

2. BATCH AND INCREMENTAL MAP ADAPTATION

2.1. MAP adaptation

The MAP framework provides a way of incorporating prior information in the estimation process, which is particularly useful for dealing with problems posed by sparse training data, for which Maximum Likelihood (ML) estimation gives inaccurate estimates. The difference between MAP and ML estimation lies in the assumption of an appropriate prior distribution of the parameters to be estimated. If λ, which is assumed to be a random vector, is the parameter vector to be estimated from the observation X

0-7803-2431-5/95 $4.00 © 1995 IEEE

with probability density function (pdf) $f(X|\lambda)$ and g is the prior pdf of λ, then the MAP estimate is defined as the maximum of the posterior pdf of λ, $g(\lambda|x) \propto f(X|\lambda)g(\lambda)$

We attempt to adapt the means and the variances of the HMM gaussians. The appropriate priors are gaussians and gamma distributions respectively, and the re-estimation formulae are given by [1]

$$m_{MAP} - m = \frac{c\sigma_0^2}{\sigma^2 + c\sigma_0^2}(m_{ML} - m)$$

$$\sigma^2_{MAP} = \frac{\beta + c/2\sigma^2_{ML}}{\alpha + c/2}$$

where m and σ^2 are the trained HMM parameters, σ_0^2 is the prior variance of m, α and β are the parameters of the gamma distribution such that $\beta/\alpha = \sigma^2$, and the subscript ML is used to denote the values estimated with a straightforward ML estimation. The priors can be estimated by considering the HMM parameters as random variables (RVs) and their values for each of the training speakers as samples of these RVs. Accordingly, they can be estimated with standard statistical procedures.

Extended MAP (EMAP:) EMAP explicitly introduces correlations to the MAP re-estimation framework described above. To describe EMAP we define $m = [m_1 \cdots m_K]^T$ as the collection of the gaussian means of a phoneme codebook, $m_0 = [m_{0_1} \cdots m_{0_K}]^T$ the SI model value of m, and we assume that the means of the codebook are correlated normal RVs with correlation matrix $S_0 = E(m - m_0)(m - m_0)^T$. Assuming that we obtain independent observations for the means of the codebook it can be shown that the re-estimation formula is given by

$$m_{MAP} - m_0 = S_0(S + CS_0)^{-1}C(m_{ML} - m_0) \qquad (1)$$

where $m_{ML} = [m_{ML_1} \cdots m_{ML_K}]^T$, $S = diag(\sigma_1^2, \cdots, \sigma_K^2)$, and $C = diag(c_1, \cdots, c_K)$ with c_i the adaptation EM count associated with mean i.

Class Tying: In a PTM HMM the triphones within a phoneme are represented as a mixture of K gaussians RVs M_i. We can accordingly define an RV M to represent the phoneme itself, as the average of the gaussians that model this phoneme, $M = \sum_i c_i M_i / \sum_i c_i$, with c_i the trained EM count of RV M_i. As a linear combination of gaussian RVs M is also a gaussian RV, and the same adaptation procedures used for the HMM gaussians. The advantage in adapting the phoneme model independently is twofold. First, there will be more samples for M than for any of the individual M_is, hence the shift of M will be more robustly estimated. As such, it can be used to smooth the shifts of the individual gaussians of the phoneme codebook associated with M. Second, when used together with EMAP, tying guarantees that all the HMM gaussians will change, even when they have no adaptation data.

2.2. The Adaptation Procedure

The adaptation procedure for supervised and unsupervised tasks goes through the following steps:

1. After an EM iteration on the adaptation data, obtain sufficient statistics for the phoneme gaussians M^k and the HMM gaussians M_i^k, $k = 1, \cdots, N_{phonemes}$.
2. Estimate the EMAP shifts for the set of phoneme gaussians M^k
3. Estimate the EMAP shifts independently for each of the phoneme codebooks
4. Propagate the phoneme gaussian shifts to the HMM gaussians:

$$\Delta m^k_{i,EMAP} = \epsilon \Delta m^k_{i,EMAP} + (1 - \epsilon)\Delta m^k_{EMAP} \qquad (2)$$

   where $\Delta m^k_{i,EMAP}$ represents the shift of a HMM gaussian and Δm^k_{EMAP} represents the shift of the phoneme model associated with the gaussian in question.
5. Re-estimate the variances around the adapted means

For batch adaptation this procedure is applied once on the enrollment data. For incremental adaptation the procedure is applied every sentence. In the latter case supervision for the EM iteration of step 1 is provided by the recognizer output either as is (unsupervised incremental adaptation) or after correcting the errors (supervised incremental adaptation).

3. INSTANTANEOUS ADAPTATION

Together with the batch and incremental adaptation we are currently exploring the feasibility of instantaneous adaptation techniques, that is unsupervised adaptation techniques that (i) require minimal amount of data – one sentence and, (ii) attempt to improve recognition on the same data that is used to estimate the adaptation transformation. We use the knowledge that this sentence comes from the same speaker and it was uttered in a roughly stationary environment. Hence we can assume that most of the difference between the speaker in question and the recognition model lies on the speaker spectral characteristics. Because the adaptation data are very limited we constrain this transformation to operate only on the phoneme level. For example, when we adapt the codebook means for instantaneous adaptation we set ϵ in (2) to zero, forcing all the gaussians of a phoneme codebook to share the shift that we estimated for the phoneme itself.

For instantaneous adaptation, in addition to EMAP re-estimation, we are using a procedure based on a similar model to that used in our previously reported Probabilistic Spectral Mapping (PSM) algorithm [4]. PSM models the difference between the speaker and the trained model spectral spaces as a probabilistic transformation of the gains of the gaussian mixtures (mixture weights). We briefly describe how we can use the EM algorithm to estimate the transformation, using a single sentence and the alignment of this sentence to the trained model. The estimation is based on the maximization of the joint likelihood of the observed test speaker speech and the unobserved data of what the model predicts a training speaker would have said.

During recognition we obtain the sequence of observations $Y = \{y_1 ... y_T\}$ and seek to find the state sequence $S = \{s_1 ... s_T\}$ that maximizes the probability of the observations. Without

[1] All of the formulae to be presented assume that the dimensionality of the observation space is 1, which of course is not true. Our system however uses diagonal covariance gaussians, and all of the problems we are dealing with are decomposed to many independent problems, one for each dimension.

adaptation we assume that the model $p(y|s)$ is the same as the trained model $p(x|s)$. PSM introduces a channel that models the difference between the training and test speakers. In the absence of the channel the model generates symbols according to $p(x|s)$. However, we observe these symbols through the speaker channel $p(y|x,s)$. So the equivalent HMM can be constructed, $p(y|s) = E_x p(y,x|s) = \sum_x p(y,x|s)p(x|s)$ If we had aligned triplets (x_t, y_t, s_t) the channel would be simply estimated as

$$p(y=j|x=i,s=k) = \frac{n_{ijk}}{\sum_l n_{ilk}}$$

where n_{ijk} is the number of observed triplets $(x_t = i, y_t = j, s_t = k)$. We can obtain these triplets only when we have a training speaker utter the same sentence as the test speaker. But we have only the test speaker sentence, so we have to represent what the training speakers might have said by a random variable $\hat{X}$. So in this case the estimation procedure involves an EM iteration to converge to a model that best characterizes the test speaker [6].

The instantaneous adaptation procedure can be summarized as follows:

1. Decode the sentence in question and obtain an alignment of frames to phonemes.
2. Estimate a phoneme dependent PSM and apply it to the model's mixture weights.
3. Estimate an EMAP shift for each of the phoneme models, then apply this shift to all gaussians associated with that phoneme.
4. Repeat the recognition.

4. EXPERIMENTAL RESULTS

To evaluate the adaptation techniques outlined in sections 2 and 3 we performed a series of experiments on various 5K word tests of the ARPA Wall Street Journal (WSJ) corpus. We used the BBN BYBLOS PTM speech recognizer that uses codebooks of 64 gaussians to model each of the 46 phonemes of the system. In all our experiments separate models were obtained for male and female speakers. These models were trained on the SI-284 portion of the corpus.

4.1. Batch Adaptation

The first experiment for batch adaptation was to measure the contribution of each of the adaptation components. This experiment was performed on the 1993-H2-dev [2] set. An adaptation model was estimated for each of the test speakers based on the 40 enrollment sentences that are provided with the test. The adapted model was then used to rescore the SI model n-best [7]. The results are summarized in Table 1. The most important contributor to the MAP gain is the adaptation of the means (12.3% reduction in error rate if used alone), then the adaptation of the variances (6.1%), and then the adaptation of the mixture weights (3.9%). We also note that EMAP adaptation is performing slightly better than simple MAP (6.28% word error rate for EMAP versus 6.47% for MAP). We did not produce the same breakdown for the non-native speakers yet. The only relative experiment we have done so far is to evaluate if the adaptation of mixture weights is helping more here than it helps for the natives, and indeed it does.

Condition	Word Error	Improvement
No adaptation	7.16%	-
MAP, mixture weights	6.88%	3.9%
MAP, variances	6.72%	6.1%
MAP, means	6.47%	9.6%
EMAP, means	6.28%	12.3%
EMAP, means + MAP for variances	6.02%	15.9%
EMAP means, + MAP for variances and mixture weights	5.96%	16.8%

Table 1: Breakdown of how much each component of the batch adaptation helped the PTM system.The test set used here is the 1993-H2-dev (native speakers). For this experiment we are simply rescoring the SI n-best

Table 2 presents the improvement in performance for various test sets when the adapted model is used during recognition (and not only n-best rescoring). We note that for native speakers (1994-S0-dev and 1994-S0-eval tests) a 23-28% reduction in word error rate is realized. For the non-native speakers (1993-S3-dev, 1994-S3-dev and 1994-S3-eval) the gain is much bigger, with the error rate being reduced by a *factor* of 2.5 to 3.5, depending on the test set. Also note that the difference in word error rate between native and non-native speakers is reduced from a factor of 2.5 to 5 before adaptation to less than a factor of 2 after adaptation, which may be quite acceptable.

Test set	Word Error		Improvement
	No adaptation	Adaptation	
1994-S0-dev	7.5%	5.4%	28%
1994-S0-eval	7.7%	5.9%	23%
1993-S3-dev	35.6%	11.0%	70%
1994-S3-dev	30.9%	10.4%	67%
1994-S3-eval	24.7%	9.8%	60%

Table 2: Performance of the PTM system with and without batch supervised adaptation on various test sets (the S0 tests are on native speakers and the S3 tests on non-natives.)

4.2. Incremental Adaptation

Similar to the batch adaptation experiments we tested our incremental adaptation algorithm on both native and non-native speakers. However, because the performance of incremental adaptation improves with the more data presented and because the length of the native and non-native speaker tests were different a straightforward comparison is not possible. It is also hard to present all the aspects of an incremental adaptation (average error rate, error rate as a function of the number of sentences seen, supervised vs.

[2] The naming convention used here for the test sets is ¡year the test was released¿-¡name of the test¿-¡purpose of the test¿. The purpose of the test can be either dev (for development) or eval (for evaluation). For example 1993-H2-dev means the development test released in 1993 for the H2 test, which is an all purpose native speaker 5k word vocabulary test.

unsupervised adaptation etc.) in the limited space of this paper. Accordingly, we will restrict ourselves to outlines of our most important conclusions instead of detailed results:

- The performance improves by 40% *after* approximately 100 sentences for the native speakers and by a factor of 2 *after* 40 sentences for non-native speakers.
- With incremental adaptation performance improves fast for the first couple dozens of sentences, but the rate of improvement is gradually slowed down. However, we did not see the improvement in performance saturate, at least for tests with a length of a 100 sentences.
- The performance of unsupervised incremental adaptation is only slightly worse than the performance of supervised incremental adaptation.
- The adapted model recognizes much faster that the SI model, actually much more than the amount needed to compensate for the computation introduced by the adaptation algorithm itself.

4.3. Instantaneous Adaptation

Table 3 presents the results we have obtained so far with instantaneous adaptation. The improvement for natives is expectedly small, since there is no significant difference between the SI model and the test speakers, at least as far as the very constrained adaptation transformation we are using can detect. For the non-native speakers, were the spectral characteristics are sufficiently different on the phoneme level, instantaneous adaptation gives a sizable gain. PSM instantaneous adaptation improved performance by 16% over no adaptation and EMAP instantaneous adaptation improved by 24%. When we used both techniques simultaneously the reduction in word error rate was 29%. Note however that the performance with instantaneous adaptation is a factor of 2.5 worse that the performance we can achieve for the same set with supervised batch adaptation on 40 enrollment sentences.

Condition	Word Error	
	1994-S0-dev (natives)	1993-s3-dev (non-natives)
No adaptation	7.7%	35.7%
PSM	7.6%	29.8%
EMAP	7.5%	27.2%
PSM+EMAP	7.5%	25.3%

Table 3: Performance of the PTM system with and without instantaneous adaptation for natives and non-native speakers.

5. DISCUSSION

We have presented a framework for MAP adaptation of large scale HMM speech recognizers. This framework, by taking advantage of correlations present among the system's parameters and by introducing various degrees of tying among them overcomes a distinct disadvantage of MAP estimation, that of changing only parameters for which adaptation data are present, keeping at the same time its asymptotic properties. We have evaluated our adaptation algorithm for batch and incremental adaptation with a Phonetically Tied Mixture (PTM) HMM and have obtained very encouraging results for both native and (especially for) non-native speakers of American English. Currently we are working on applying the adaptation algorithm presented in more complex continuous mixture HMM, were we have to adapt orders of magnitude more parameters than the PTM system.

We have also touched on the issue of instantaneous adaptation. We have shown that when there is a big difference between the test speaker and the SI model used for recognition, a sizable gain can be achieved by adapting the model based on a single sentence and then repeating the recognition on the same sentence. It will be interesting to see how instantaneous adaptation techniques work together with incremental adaptation. We are also still looking for a more appropriate set of features to adapt with instantaneous adaptation, features that are few enough to be estimated robustly form a single (and usually misrecognized) sentence but yet powerful enough to capture the different characteristics of a new speaker.

6. ACKNOWLEDGMENT

This work was supported by the Advanced Research Projects Agency and monitored by the Office of Naval Research under Contract No. N00014-91-C-0115, and by Ft. Huachuca under Contract Nos. DABT63-94-C-0061 and DABT63-94-C-0063 . The content of the information does not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred.

REFERENCES

[1] J.L. Gauvain and C.H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Observations of Markov Chains", *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 2, pp. 291–298, April 1994.

[2] C.H. Lee, C.H. Lin and B.H. Juang, "A Study in Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", *IEEE Trans. Sig. Proc.*, Vol. 39, No. 4, pp. 806–814, April 1991.

[3] M.J. Lasry and R.M Stern, "A posteriori Estimation of Correlated Jointly Gaussian Mean Vectors", *IEEE Int. Transactions. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 4, July 1984, pp. 530–535.

[4] R. Schwartz, Y.L. Chow and F. Kubala, "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping", IEEE ICASSP-87, Apr. 1987, paper 15.3.1.

[5] Paul, D., J. Baker, "The Design for the Wall Street Journal-based CSR Corpus", *Proc. of DARPA Speech and Natural Language Workshop*, Harriman, NY, February 1992, pp. 357-362.

[6] J.R. Rohlicek, V. Beattie, personal communication.

[7] R. Schwartz, S. Austin, F. kubala, J. Makhoul, L. Nguyen, P. Placeway and G. Zavaliagkos, "New uses for the N-best Sentence Hypotheses Within the BYBLOS Speech Recognition System",*IEEE Int. Conf. Acoustics, Speech and Signal Processing* , San Fransisco, March 1992, pp 1-4.