HMM TRAINING BASED ON CV-EM AND CV GAUSSIAN MIXTURE OPTIMIZATION

Takahiro Shinozaki[*], *Tatsuya Kawahara*

Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

[*]staka@ar.media.kyoto-u.ac.jp

ABSTRACT

A combination of the cross-validation EM (CV-EM) algorithm and the cross-validation (CV) Gaussian mixture optimization method is explored. CV-EM and CV Gaussian mixture optimization are our previously proposed training algorithms that use CV likelihood instead of the conventional training set likelihood for robust model estimation. Since CV-EM is a parameter optimization method and CV Gaussian mixture optimization is a structure optimization algorithm, these methods can be combined. Large vocabulary speech recognition experiments are performed on oral presentations. It is shown that both CV-EM and CV Gaussian mixture optimization give lower word error rates than the conventional EM, and their combination is effective to further reduce the word error rate.

Index Terms**—** HMM, Gaussian mixture, cross-validation, parameter estimation, structure optimization

1. INTRODUCTION

In order to obtain high recognition performance, precise modeling of speech sounds is important. In general, a probabilistic model can express a more complex distribution when it has a larger number of parameters. Thus, it is necessary for acoustic models used for large vocabulary continuous speech recognition (LVCSR) to have many parameters. On the other hand, the model parameters need to be estimated before they are used in speech recognition. When many parameters are estimated from a limited amount of data, the estimation involves errors and the error degrades the model performance. In other words, the model over-fits to the given training data and loses the ability to generalize. Basically, the error becomes larger when the amount of data is small relative to the number of model parameters. However, the amount of the estimation error also depends on the training algorithm. To improve speech recognition performance, it is important to develop a training algorithm that is able to accurately estimate large models from a limited amount of data as well as optimizing the model size.

Acoustic models for LVCSR are usually implemented as Gaussian mixture HMMs. A general recipe of training Gaussian mixture HMMs is to start with a single Gaussian HMM and then repeat the expectation maximization (EM) algorithm along with the mixture splittings. The problems of this procedure are that EM training is susceptible to over-training and there is no mechanism to find the best mixture size. Moreover, the model training by the EM algorithm can be even instable. For example, when training a two-mixture Gaussian distribution, the algorithm sometimes produces a mixture distribution in which one of the Gaussians covers only a few data points with very small variance and the other Gaussian spans the rest of the data points. Obviously, such a model is not desirable in terms of generality.

These problems originated from using training set likelihood as an objective function for the parameter estimation. Because the likelihood is evaluated by a model whose parameters are estimated on the same data, it is positively biased. The bias becomes especially large when the amount of training data is small relative to the number of model parameters. In order to address these problems by removing the bias, we have proposed cross-validation EM (CV-EM) algorithm [1] and cross-validation (CV) Gaussian mixture optimization method [2] that replace the conventional training set likelihood with CV likelihood. While CV-EM uses the CV likelihood to estimate the expected sufficient statistics, CV Gaussian mixture optimization uses the CV likelihood to select a pair of Gaussian components that should be merged.

In the previous studies, these CV training methods were proposed and evaluated on the different tasks. Here, we evaluate these algorithms and their combination on the same large vocabulary speech recognition task.

This paper is organized as follows. The CV based training algorithms are briefly reviewed in Section 2. Experimental conditions are shown in Section 3 and the results are presented in Section 4. Finally, a summary and future works are given in Section 5.

2. CV TRAINING ALGORITHMS

In this section, the CV-EM algorithm and the CV Gaussian mixture optimization method are briefly reviewed. One of the novel points of these algorithms is the efficient evaluation of the CV likelihood by utilizing sufficient statistics. The details of these algorithms can be found in the original papers [1, 2].

978-1-4244-1746-9/07/$25.00 ©2007 IEEE

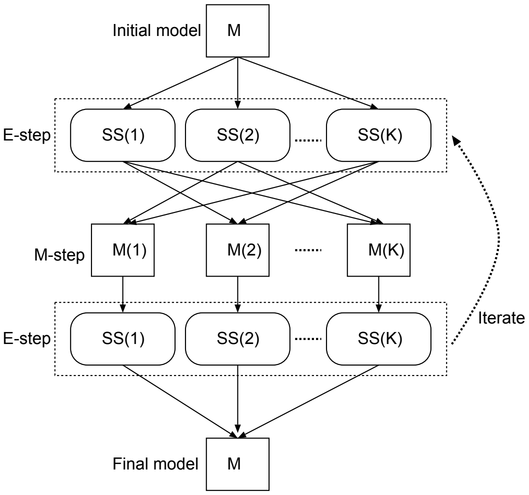

Fig. 1. *CV-EM training. SS(i) is a sufficient statistics estimated on the i-th data subset. M(i) denotes the i-th CV model estimated without using the i-th data subset.*

2.1. CV EM algorithm

CV-EM introduces CV into the framework of the EM algorithm that repeats the expectation step (E-step) and the maximization step (M-step) alternatively. In the E-step of the EM algorithm, expected sufficient statistics are estimated given a current model, and in the M-step, the model parameters are updated by the maximum likelihood criterion based on the sufficient statistics. Because the E-step and the M-step use the same training data, the EM iteration reinforces the bias for particular training samples and is susceptible to over-fitting. The key idea behind the CV-EM algorithm is to separate data used in the E-step and the M-step. Since there is no overlap in the data used for the E-step and the M-step, the potential for over-fitting is reduced.

Figure 1 shows the training procedure of CV-EM. The training data is partitioned into K subsets. In the E-step, sufficient statistics are independently calculated for each subset. Then, in the M-step, K CV models are estimated by accumulating all but one sufficient statistics. Each CV model is used in the next E-step to estimate the new sufficient statistics for the data subset that has been excluded from the parameter estimation of that model. Thus, there is no overlap in data used in the E-step and the M-step. The E-step and the M-step are repeated as in conventional EM training and the final model is obtained by merging all the sufficient statistics. When the training data size is large compared to K, CV-EM has the same order of computational cost as the EM algorithm.

Figure 2 compares test set likelihood of Gaussian mixture models trained by the EM and the CV-EM algorithms. The training and test data were sampled from 4-dimensional 8-mixture diagonal Gaussian distributions whose parameters were randomly defined. Because of the over-fitting problem, the test set likelihood does not monotonically increase for the training iterations. It decreases for larger training iterations especially when a small training set is used compared to the model size. CV-EM is more robust against the over-fitting problem than EM. Therefore, CV-EM can accurately estimate larger model than conventional EM for a given amount of training set.

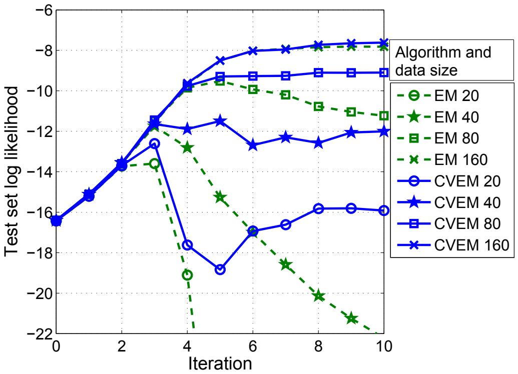

Fig. 2. *Test set likelihood of GMMs trained by EM and CV-EM with varying training set sizes.*

2.2. CV mixture optimization algorithm

Given a model with large mixtures, a strategy to optimize Gaussian mixture distribution is to select and merge a pair of components based on an objective function step by step until a termination criterion is satisfied. Figure 3 shows an example of the merging process. The most popular choice for the objective function is the training set likelihood. However, the drawback with using the likelihood is that it is optimistically biased and not reliable. Related to the bias, the likelihood always decreases for the component merging and it is difficult to know when to stop the merging process.

CV mixture optimization algorithm uses CV likelihood instead of the conventional training set likelihood. The CV likelihood is less biased and is more reliable than the training set likelihood. Therefore, the CV likelihood behaves as the likelihood estimated on the new data. The optimal point for the Gaussian merging iterations can be found as the maximum point of the CV likelihood.

By utilizing pre-computed sufficient statistics, the CV likelihood can be efficiently evaluated for all combinations of the components. Because the algorithm is based on the data-driven method, it is expected to be more robust than the information theoretic model selection criteria such as minimum description length (MDL) criterion, which often requires an empirical tuning factor to compensate for errors in the theoretical assumptions [3].

Fig. 4 shows an example of the likelihood that is estimated during the Gaussian merging optimization for a cer-

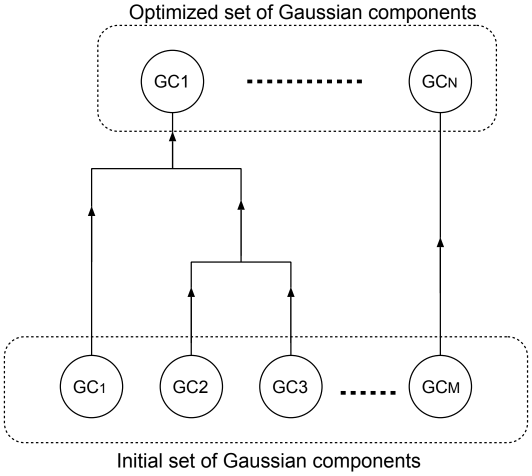
Optimized set of Gaussian components
GC1
GCN
GC1
GC2
GC3
GCM
Initial set of Gaussian components

Fig. 3. *Gaussian mixture optimization.*

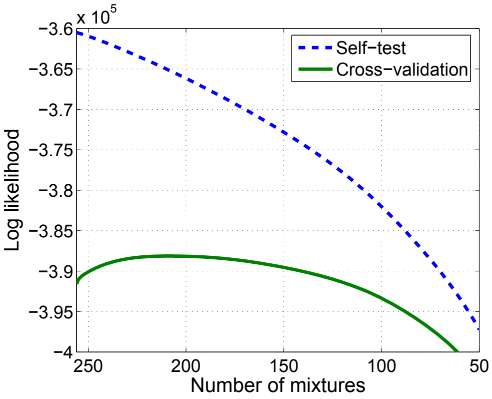

Fig. 4. *Gaussian component merging and GMM likelihood evaluated on the training set. "Self-test" is the conventional training set likelihood and "cross-validation"is the likelihood by the CV method. The conventional likelihood takes a larger value than the CV likelihood because of the positive bias.*

tain HMM state with 256 mixtures. Conventional likelihood takes a larger value than the CV likelihood because of the optimistic bias. The increase of the CV likelihood indicates that the model generality is improved by merging the components and the decrease indicates that the model is becoming too small. Therefore, the optimal number of mixtures is around 210 in this case.

3. TRAINING PARADIGM AND EXPERIMENTAL SETUPS

HMM acoustic models were trained using the CV training algorithms independently or in combination. When the Gaussian mixture optimization is performed, there are several possibilities of how to apply it. For example, it can be applied only once using an HMM with large mixtures as an input model. A problem with this strategy is that it is not obvious how to choose the number of mixtures for the initial model. The other strategy is to repeat the merging process along with mixture splitting. In this way, the initial mixture size problem is avoided. In addition, a positive effect is expected in finding better local optima as it kneads the mixtures by repeatedly absorbing unnecessary components and increasing the survived Gaussians. In this work, the latter training procedure is adopted. The HMMs were trained with the following procedure:

1. Input 1-mixture tied-state HMM as an initial model.
2. Randomize and uniformly partition the training data.
3. Iterate EM or CV-EM for five times.
4. Optimize Gaussian mixtures by merging components. Either the CV mixture optimization method or the MDL criterion based method is used. Output HMM.
5. Split and double the number of the mixtures by duplicating the parameters with small deviation. Go to step 2.

In addition to performing the mixture optimization using the cross-validation based method, MDL criterion based optimization was also investigated. The tuning factor for the MDL criterion based method was set to 1.0 based on preliminary experiments in which 256 mixture Gaussian HMMs were optimized with different tuning factors and evaluated for the test set.

In the following, we count step 2 through step 5 as one training iteration. The random partitioning was performed for each training iteration. If the Gaussian merging in step 4 is not performed, then the number of Gaussians in the HMM is simply doubled for each training iteration. We refer to this procedure with the EM training as a baseline.

The Gaussian mixture HMMs were tied-state model with 1000 states. They were trained from 30 hours of a subset of the Corpus of Spontaneous Japanese (CSJ) [4]. The utterances were from academic presentations. Feature vectors had 39 elements comprising of 12 MFCC and log energy, their delta, and delta delta. The HTK toolkit [5] was used for the EM training. In order to support the operations on sufficient statistics, a modified version of HTK was used for the CV-EM. The language model was a trigram model trained from 6.8M words of academic and extemporaneous presentations from the CSJ. Test set was the CSJ evaluation set that consisted of 10 academic presentations given by male speakers. Speech recognition was performed using the Julius decoder [6]. The number of the CV folds was 30 for both CV-EM and CV Gaussian mixture optimization.

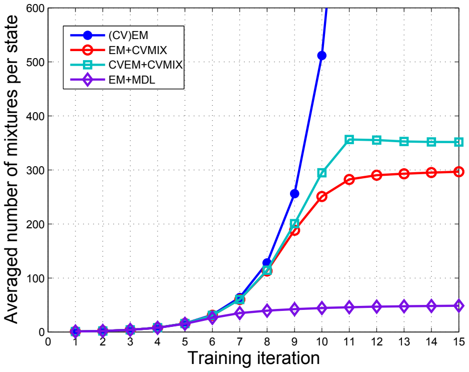

Fig. 5. *Number of training iterations and averaged number of mixtures per state. "(CV)EM" shows the number of mixtures when the mixture optimization is not performed. "EM+CVMIX" and "CVEM+CVMIX" are the results when CV mixture optimization is combined with EM and CV-EM, respectively. "EM+MDL" is the combination of EM and MDL based mixture optimization.*

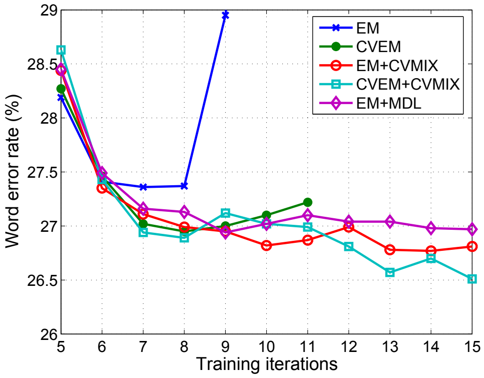

Fig. 6. *Number of training iterations and test set word error rate. "EM" and "CVEM" are the results of EM and CV-EM without the mixture optimization. "EM+CVMIX" is the combination of the CV mixture optimization with EM, and "CVEM+CVMIX" is the combination with CVEM. "EM+MDL" is the combination of MDL based mixture optimization with EM.*

4. EXPERIMENTAL RESULTS

Fig. 5 plots the averaged number of mixtures per state for the training iteration. When the merging optimization was performed, the number of mixtures first increased exponentially and then gradually converged to a constant value. This is because the Gaussian merging hardly occurs when the model is small. As the number of mixtures increased, the merging process effectively started to work. After sufficient iterations, the number of merged components became equal to the number of splits and a balance in the total number of mixtures was reached. The Gaussian merging optimization gave larger mixture size when it is combined with CV-EM than when combined with EM. This is probably because CV-EM makes better use of more parameters. The MDL criterion based mixture optimization gave smaller model sizes than the CV based optimization.

Fig.6 shows word error rates by the models trained by the EM and the CV-EM algorithms with and without the mixture optimization. When EM was used without the mixture optimization, the lowest word error rate of 27.4% was obtained at the seventh iteration and then the performance began to decrease for the training iterations. This is because the sparseness problem arises as the model size gets large. CV-EM always gave similar or lower word error rates than EM. Especially, it was much more robust than EM for the larger model sizes. The lowest word error rate of 27.0% was obtained at eighth iteration.

When EM was combined with the CV mixture optimization, the model size was automatically controlled and the sparseness problem was mostly avoided. Moreover, as the iterations proceeded, it gave lower word error rates than EM by finding better local optima.

The combination of EM and MDL based mixture optimization gave lower word error rates than EM, but higher than the combination of EM and the CV based mixture optimization. The lowest word error rate was 26.9%.

Among the training strategies, the lowest word error rate of 26.5% was obtained by the combination of CV-EM and the CV mixture optimization with 15 training iterations. The relative reduction of the error rate was 3.3% compared to the lowest error rate of the EM training.

5. SUMMARY AND FUTURE WORKS

In this study, we have evaluated the CV-EM algorithm, the CV based mixture optimization method, and their combination. It has been confirmed that both CV-EM and CV based mixture optimization method gave lower word error rates than the conventional EM. Moreover, the recognition performance was further improved by combining the two CV training schemes. These results indicate that the CV likelihood based training methods are better able to train precise models than conventional training algorithms from a limited amount of training data without suffering the over-fitting problem.

Future works include the combination with the decision tree clustering method based on the CV likelihood [7]. By combining these algorithms, all the likelihoods used in the basic HMM training procedure are substituted by the CV likelihood. This will make the training process more robust to over-fitting and higher recognition performance is expected.

6. ACKNOWLEDGMENTS

This study was supported by the Special Coordination Funds for Promoting Science and Technology, Ministry of Education, Culture, Sports, Science and Technology, Japan.

7. REFERENCES

[1] T. Shinozaki and M. Ostendorf, "Cross-validation and aggregated EM training for robust parameter estimation," *Computer speech and language*, accepted.

[2] T. Shinozaki and T. Kawahara, "Gaussian mixture optimization for HMM based on efficient cross-validation," in *Proc. Interspeech*, 2007, accepted.

[3] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech*, 1997, vol. 1, pp. 99–102.

[4] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," in *Proc. SSPR2003*, 2003, pp. 135–138.

[5] S. Young *et al.*, *The HTK Book*, Cambridge University Engineering Department, 2005.

[6] A. Lee, T. Kawahara, and S. Doshita, "An efficient two-pass search algorithm using word trellis index," in *Proc. ICSLP*, 1998, pp. 1831–1834.

[7] T. Shinozaki, "HMM state clustering based on efficient cross-validation," in *Proc. ICASSP*, Toulouse, 2006, vol. I, pp. 1157–1160.