Discriminative Splitting of Gaussian/Log-Linear Mixture HMMs for Speech Recognition

Muhammad Ali Tahir, Ralf Schlüter, Hermann Ney

Lehrstuhl für Informatik 6, Computer Science Department
RWTH Aachen University, Aachen, Germany
{tahir,schlueter,ney}@cs.rwth-aachen.de

***Abstract*—This paper presents a method to incorporate mixture density splitting into the acoustic model discriminative log-linear training. The standard method is to obtain a high resolution model by maximum likelihood training and density splitting, and then further training this model discriminatively. For a single Gaussian density per state the log-linear MMI optimization is a global maximum problem, and by further splitting and discriminative training of this model we can get a higher complexity model. The mixture training is not a global maximum problem, nevertheless experimentally we achieve large gains in the objective function and corresponding moderate gains in the word error rate on a large vocabulary corpus**

I. Introduction

In a typical speech recognition system, the feature extraction phase is modeled such that it should only retain that part of the audio signal that is relevant to the phonetic content. The irrelevant information i.e. speaker specific and environmental content should be removed as much as possible. Popular feature extraction schemes like the Mel frequency cepstral coefficients (MFCC) do capture sufficient phonetic information, yet they still contain a lot of extra information. A proof of this is the fact that MFCC are also used for speaker identification tasks [1]. To model this variability we use a mixture of Gaussians to represent each phone class (usually a triphone state in a large vocabulary speech recognition system). In a generative maximum likelihood hidden Markov model based system, a single Gaussian is trained for each triphone state, and then it is iteratively split into a large number of Gaussians, to better fit the training data. The details of such a process can be found in [2].

Discriminative training [3] of mixture models has generally resulted in better word error rates (WER) than the conventional ML training [2]. Contrary to ML, the discriminative training strives to maximize the separation between different classes, so that they are more readily distinguishable. In this case the actual likelihood of the data is not important. However, experiments show that the gains due to discriminative training are particularly high for low complexity models i.e. a small number of Gaussians per triphone state. For a large number of Gaussians per state the performance of a discriminatively trained system is only slightly better than a ML system. The discriminative criteria like maximum mutual information (MMI) are optimized by iterative methods like gradient descent or extended Baum-Welch, which only guarantee a local optimum.

The Gaussian parameter splitting may also be accomplished discriminatively to obtain better fitting models, as in [4] where results on a digit recognition task are presented. The emphasis there is to retain the performance of a good system while successively reducing the number of parameters. In [5], a measure of classification error is used to determine the candidate densities to be split. In [6], the mixture densities are split discriminatively, and then further trained by ML estimation.

In this work we shall follow a consistently discriminative path as much as possible; using a simple single density acoustic model from a ML estimate, followed by a combined training and splitting, both done discriminatively while optimizing the MMI objective function. Our emphasis is to train a large vocabulary system with several million parameters. We use log-linear discriminative training as explained in section II-A, because it guarantees a global maximum for a single density per triphone state.

The rest of this paper is organized as follows. Section II introduces the conversion of Gaussian mixture models to log-linear form, and the discriminative training procedure. Section III describes discriminative splitting. In Section IV, experiments and their results on a large vocabulary corpus are presented. Finally, Section V provides the conclusion and future outlook.

II. Log-linear Mixture Model

In [7] it has been shown that the posterior form of Gaussian HMM can be represented as a heteroscedastic conditional random field. This simplifies to a conditional random field (CRF) or log-linear model for the case of a pooled covariance HMM. The optimization of a log-linear model is a convex problem according to the maximum entropy principle [8]. For a fixed alignment between the feature vectors and the HMM states, and a single density per state, the corresponding log-linear model has a global maximum, that can be reached regardless of the intial values of parameters. This has also been shown experimentally in [9]. Another similar work is [10] although it assumes a different structure of the HMM. A

useful property of the log-linear models is that they can be used to combine features from different knowledge sources [11], as the optimization is robust to feature scaling and linear dependencies between different features.

Let the speech feature vectors x_1^T belong to one of $s = 1, ..., S$ generalized triphone classes, derived from classification and regression trees (CART); each class with Gaussian parameter set $\theta_s = \{\mu_s, \Sigma_s\}$. After expanding $p_\theta(x|s)$ in its Gaussian form and collecting the terms of x, the posterior probability becomes

$$\begin{aligned} p_\theta(s|x) &= \frac{p(s)p_\theta(x|s)}{\sum_{s'} p(s')p_\theta(x|s')} \\ &= \frac{\exp(x^\top \Lambda_s x + \lambda_s^\top x + \alpha_s)}{\sum_{s'} \exp(x^\top \Lambda_{s'} x + \lambda_{s'}^\top x + \alpha_{s'})} \end{aligned} \tag{1}$$

In Equation 1, the new parameters $\Lambda_s \in \mathbf{R}^{\mathbf{D}\times\mathbf{D}}$, $\lambda_s \in \mathbf{R}^{\mathbf{D}}$ and $\alpha_s \in \mathbf{R}$ are present in log-quadratic form. Note that here the posterior probability is directly modeled. The numerator is not normalized and therefore does not conform to the constraints of a probability distribution.

A pooled covariance matrix Σ leads to

$$p_\theta(s|x) = \frac{\exp(\lambda_s^\top x + \alpha_s)}{\sum_{s'} \exp(\lambda_{s'}^\top x + \alpha_{s'})} \tag{2}$$

which is log-linear with respect to the parameter x.

In case of mixture densities, a hidden variable for the mixture components need to be introduced. The corresponding posterior probability is

$$p_\theta(s|x) = \frac{\sum_l \exp(\lambda_{s,l}^\top x + \alpha_{s,l})}{\sum_{s',l} \exp(\lambda_{s',l}^\top x + \alpha_{s',l})} \tag{3}$$

for $l = 1...L_s$ mixture parameters in each class s.

A. Discriminative Training of Log-Linear Parameters

The frame level objective function is

$$\begin{aligned} \mathcal{F}^{(frame)}(\Lambda) &= -\tau_\Lambda ||\Lambda||^2 \\ &\quad + \sum_{r=1}^{R} \sum_{t=1}^{T_r} w_s \log p_\Lambda(s_t|x_t) \end{aligned} \tag{4}$$

$$p_\Lambda(s_t|x_t) = \frac{\exp\left(\lambda_{s_t}^\top x_t + \hat{\alpha}_{s_t}\right)}{\sum_{s'} \exp\left(\lambda_{s'}^\top x_t + \hat{\alpha}_{s'}\right)} \tag{5}$$

for a fixed alignment s_1^T where the state parameters are $\Lambda_s = \{\lambda_s, \alpha_s\}$. τ_Λ is the regularization parameter to increase robustness and avoid over-fitting. w_s are state weights which could be tuned to give less weight to some states e.g. silence which occupies a large number of states in the alignment. $\hat{\alpha}_s = \alpha_s + \log p(s)$, $p(s)$ is the prior probability of state s and R is the total number of sentences in the training corpus. The state priors are later subtracted from $\hat{\alpha}_s$ for recognition, because for recognition we use language model priors instead of state priors. The objective function is frame-level Maximum Mutual Information (MMI), with an extra regularization term. The MMI optimization may also be done at sentence level, by using language-model probabilities as priors.

B. Optimization Procedure

For the optimization of the objective function in Equation 4 we use the general purpose RPROP algorithm [12]. RPROP is a first order optimization algorithm that takes only the sign of the partial derivatives into account. The weights for parameters are increased if there was no sign change in the partial derivatives in the last iteration, and vice versa. In all the following experiments in Section IV the RPROP algorithm is used for optimization.

The MMI optimization on a large training corpus can be computationally expensive. A remedy is to use the Viterbi approximation for the optimization of mixture densities. This means for each $p(x|s)$ using the score of the highest scoring density instead of the sum of all the densities. In practice it was found to be detrimental for the optimization process. When the Viterbi option is enabled, only those feature vectors contribute to the partial derivatives of $\lambda_{s,l}$ which lie closer to it than all other $\lambda_{s,l'}$. Therefore if a particular $\lambda_{s,l}$ strays away from the solution due to a large step size, it will not be brought back towards the solution because there are no feature vectors to contribute towards its partial derivatives. This leads to discontinuities in the partial derivatives. For this reason we calculate the sum of all the densities where possible. However, for a very large number of $\lambda_{s,l}$ parameters per state, calculating the sum does not remain feasible due to its computational requirement. Therefore in that case we have to resort to the maximum approximation. With proper limiting of the RPROP step sizes to increase its robustness, it can also give reasonable gains in the objective function.

III. Discriminative Splitting

The log-linear training is only convex for a single density per state s. For mixture density training this presents challenges as the initial guess is very important and can influence the final results for the objective function and WER. Therefore we need a method to specify a better initial guess to the training of mixture densities, so that the WER is at least as good as the word error rate of a similar but less complex model. To solve this problem we adopt an approach similar to the iterative density splitting algorithm used in a maximum likelihood framework, but applied to the log-linear parameters $\lambda_{s,l}$ instead of the means, as in the Gaussian mixtures case. All the $\lambda_{s,l}$ in state s are duplicated and a small offset is added to both new lambdas to pull them apart. The log-linear model is covariance normalized), therefore the direction of the offset is not important. Subsequent training of this newly split model causes an increase in the objective function as the new lambdas discriminatively adapt themselves to the training data.

IV. Experiments and Results

A. Speech Corpus and Baseline System

For the performance analysis of discriminative splitting, the large vocabulary continuous speech recognition task European

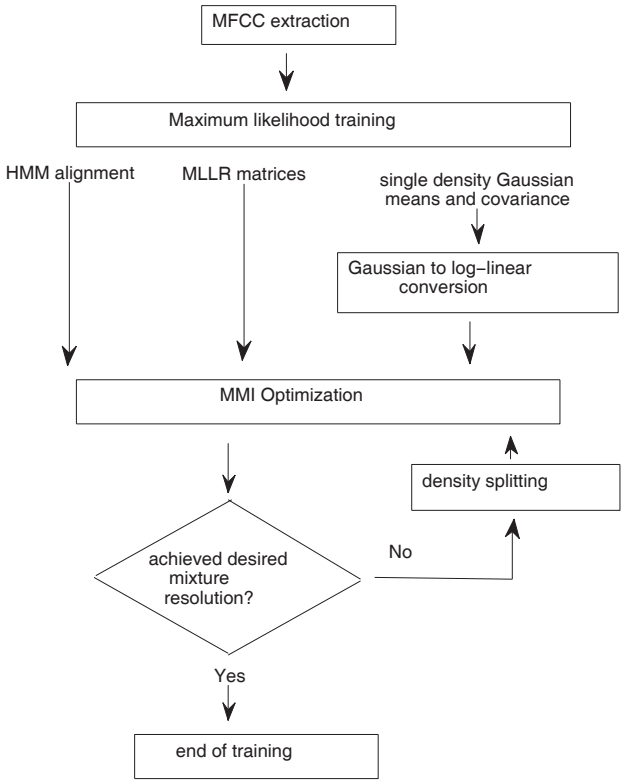

Fig. 1. *Flow diagram of the combined discriminative training and splitting process*

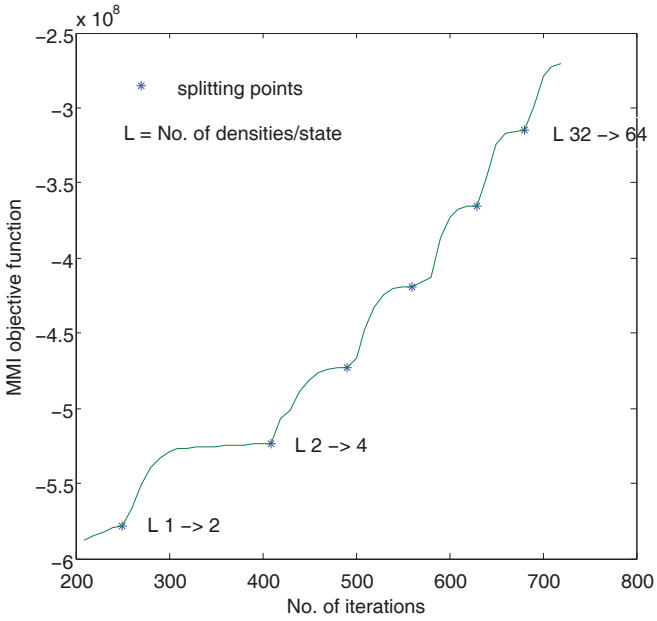

Fig. 2. *EPPS: Ascent of MMI objective function versus number of training iterations. The density splitting events are marked by **

Parliament Plenary Sessions (EPPS) is used. It is a part of 2006 TC-STAR ASR evaluation campaign. It is composed of recorded speeches of the European Parliament in British English under clean conditions. The training corpus is 40.8 hours and the evaluation corpus is 3.5 hours. The newer versions of the EPPS English corpus contain more than 100 hours of training data.

The acoustic model of the baseline system uses cross-word triphones. The lexicon contains 54k words and a trigram language model is used. The initial features are 16 MFCC features and with one energy and one voicedness feature. Nine such consecutive frames are concatenated together, and then projected by a classical LDA matrix to 45 dimensions. The classes are 4501 triphone CART leaves and a pooled covariance is used.

B. Log-linear Training

A flow diagram of the training process is shown in Fig. 1. The initial alignment between the training acoustic data and its transcription is obtained by training a Gaussian generative ML system with 256 densities per triphone state. This alignment is kept fixed during the later stage of discriminative training, as it was experimentally found that updating it had virtually no effect on the optimization procedure.

The single density Gaussian acoustic model is initialized by maximum likelihood training. This model is trained log-linearly to optimize the MMI frame-level criterion. While this is not the best criterion in terms of WER performance (sentence level MMI and MPE give better WERs), we choose frame-level MMI because it guarantees a global maximum for single densities. Once the single density optimization has converged, we split it and hence double the number of parameters. When this in turn has converged, we split it again. This process is repeated until we get 64 densities per triphone state. During the course of this process a steady increase in the objective function value is observed. For up to 8 densities per state we use a full sum of all the mixture parameters $\lambda_{s,l}$. Since the computation time doubles by doubling the resolution, therefore for 16 densities it becomes prohibitive i.e. 20 hours for a single iteration. So from here onwards we switch to viterbi optimization, and set limits on the step sizes of the RPROP procedure to increase its robustness.

Fig. 2 illustrates the progress of the objective function against the number of iterations. The blue * marked on the figure represent the points where splitting has been done and consequently the number of λ has doubled. The graph shows a consistent gain in the objective function, even for a large number of densities per state! Looking at the graph it seems that the trend would hold if we further continue splitting the log-linear models.

C. Integration of SAT MLLR and cMLLR

The baseline recognizer has another version with speaker adaptive training (SAT) with maximum likelihood linear re-

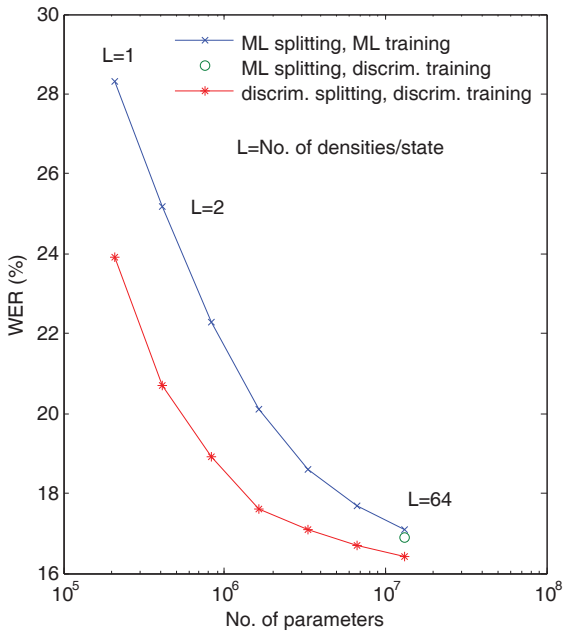

Fig. 3. *EPPS: Comparison of WER of discriminatively split and ML split log-linear models*

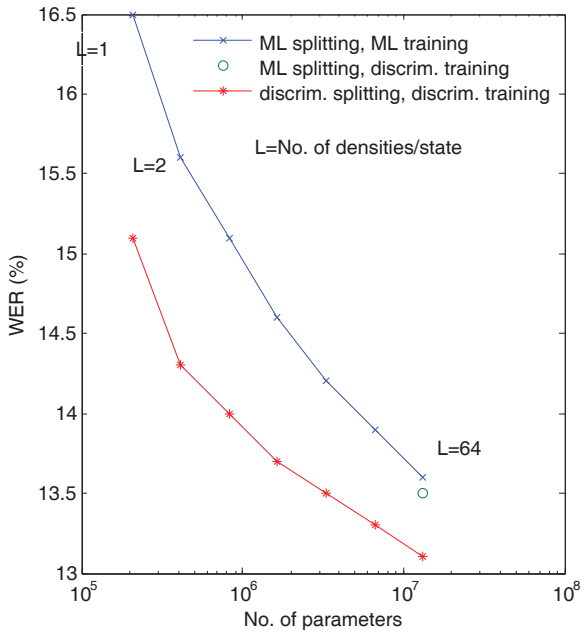

Fig. 4. *EPPS: Comparison of WER of discriminatively split and ML split log-linear models, with SAT MLLR and cMLLR*

gression (MLLR) and constrained MLLR (cMLLR). The MLLR is a feature linear transform while cMLLR transforms the parameters of the Gaussian model. Their purpose is to remove the speaker specific information. For the EPPS task SAT gives a WER improvement of 3% absolute. Therefore it should be helpful to integrate SAT into the log-linear discriminative training framework. For this purpose a maximum likelihood SAT MLLR is performed on the training data to obtain speaker specific transformation matrices. These matrices are added to the log-linear training pipeline of section IV-B and the rest of the procedure stays the same. For recognition these log-linear mixture models are converted back into Gaussian models as in [7] and then SAT MLLR and cMLLR is performed. The conversion to Gaussian mixture models is necessary since cMLLR operates on means and covariances and therefore requires a Gaussian form of the model.

D. Recognition Results

As shown in Fig. 3 and Fig. 4 , the WER differences between the single density maximum likelihood and discriminative training are quite large. However, as the number of densities increases, the difference between both is reduced. For 64 densities per state this difference is 0.7 % absolute without SAT and 0.5 % with SAT, small but still significant in relative terms.

To test the effectiveness of discriminative splitting, we take a ML Gaussian model already split as 64 densities per state, and train it discriminatively. This is a model where only the training of $\lambda_{s,l}$ is done discriminatively, and no splitting is done in between. This way we only get a 0.2 % improvement over the ML model without SAT and 0.1 % improvement with SAT. This improvement is significantly smaller than what was obtained by an integrated splitting and training. The possible reason for this could be the higher susceptibility of such an approach to get stuck in a local maximum.

An important point to note here is that the frame-level MMI criterion is not the best criterion in terms of WER. The purpose of using it for our experiments was its robustness and global maximum property (for single densities). What remains to be seen is whether the WER improvements obtained by discriminative splitting also hold for more complex criteria like MMI and MPE, and do they lead to better WER than frame-level MMI?

V. Conclusion

In this paper a technique for discriminative splitting for log-linear mixture densities is presented. For this purpose a Gaussian acoustic model is converted to log-linear form and then trained to maximize the frame-level MMI objective function. Experiments have been performed on the large vocabulary EPPS English task. The objective function gains using this approach are large and at recognition time this translates to moderate but significant gains in WER. The approach proves to be superior to another model where the Gaussian models are trained discriminatively but the splitting is done beforehand in a ML framework. However, this has only been observed on the frame-level MMI criterion. Further work in this direction would try to achieve WER gains while integrating such a

discriminative splitting approach into sentence-level MMI and MPE optimization procedures.

ACKNOWLEDGMENT

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the OSEO.

REFERENCES

[1] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining mfcc and phase information in noisy environments," in *Proc. IEEE ICASSP*, Dallas, USA, 2010, pp. 4502–4505.

[2] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, united states edition, 1993.

[3] K. Vertanen, "An overview of discriminative training for speech recognition," Tech. Rep., 2008.

[4] Y. Normandin, "Optimal splitting of hmm gaussian mixture components with mmie training," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, may 1995, pp. 449 –452 vol.1.

[5] V. Valtchev, J. J. Odell, P. Woodland, and S. Young, "Mmie training of large vocabulary recognition systems," 1997.

[6] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "A combined maximum mutual information and maximum likelihood approach for mixture density splitting," in *European Conference on Speech Communication and Technology*, vol. 4, Budapest, Hungary, Sep. 1999, pp. 1715–1718.

[7] G. Heigold, P. Lehnen, R. Schlüter, and H. Ney, "On the equivalence of Gaussian and log-linear HMMs," in *Proc. INTERSPEECH'08*, Brisbane, Australia, Sep. 2008.

[8] J. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Annals of Mathematical Statistics*, vol. 43, pp. 1470–1480, 1972.

[9] G. Heigold, D. Rybach, R. Schlüter, and H. Ney, "Investigations on convex optimization using log-linear HMMs for digit string recognition," in *Proc. INTERSPEECH'09*, Brighton, U.K., Sep. 2009.

[10] H.-K. J. Kuo and Y. Gao, "Maximum entropy direct models for speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 873 – 881, May 2006.

[11] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros, "Crf-based combination of contextual features to improve a posteriori word-level confidence measures," in *Proc. INTERSPEECH'10*, Makuhari, Japan, September 2010, pp. 1942–1945.

[12] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. ICNN'93*, San Francisco, USA, 1993, pp. 586–591.