Simultaneous Discriminative Training and Mixture Splitting of HMMs for Speech Recognition

Muhammad Ali Tahir, Markus Nussbaum-Thom, Ralf Schlüter, Hermann Ney

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

Abstract

A method is proposed to incorporate mixture density splitting into the acoustic model discriminative training for speech recognition. The standard method is to obtain a high resolution acoustic model by maximum likelihood training and density splitting, and then improving this model by discriminative training. We choose a log-linear form of acoustic model because for a single Gaussian density per triphone state the log-linear MMI optimization is a convex optimization problem, and by further splitting and discriminative training of this model we can get a higher complexity model. Previously it was shown that we achieve large gains in the objective function and corresponding moderate gains in the word error rate on a large vocabulary corpus. This paper incorporates the state of the art minimum phone error training criterion into the framework, and shows that after discriminative splitting, a subsequent log-linear MPE training achieves better results than Gaussian mixture model MPE optimization alone.

Index Terms: speech recognition, log linear modelling, discriminative training

1. Introduction

The typical speech recognition feature extraction techniques like the Mel frequency cepstral coefficients (MFCC) work well in practice, yet they still contain a lot of extra information. 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 [1].

Discriminative training [2] strives to maximize the separation between different classes, so that they are more readily distinguishable. It has generally resulted in better word error rates (WER) than the conventional ML training [1] 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 an 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.

Apart from the discriminative training of Gaussian HMM parameters, the Gaussian parameter splitting may also be accomplished discriminatively to obtain better fitting models, as in [3] 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 [4], a measure of classification error is used to determine the candidate densities to be split. In [5], the mixture densities are split discriminatively, and then further trained by ML estimation.

In this work we start from a simple single density acoustic model from an ML estimate, followed by a combined training and splitting, both done discriminatively. Our emphasis is to train a large vocabulary system with several million parameters. We use log-linear discriminative training, because it guarantees a global maximum for a single density per triphone state. The rest of this paper is organized as follows. Section 2 introduces the conversion of Gaussian mixture models to log-linear form, and the discriminative training procedure. Section 3 briefly describes discriminative splitting. In Section 4, experiments and their results on a large vocabulary corpus are presented. Finally, Section 5 provides the conclusion and future outlook.

2. Log-linear Mixture Model

In [6] it has been shown that the posterior form of a Gaussian HMM can be represented as a heteroscedastic conditional random field. This simplifies to a conditional random field (CRF) or a 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 [7]. 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 [8]. Another similar work is [9] although it assumes a different structure of the HMM. A useful property of log-linear models is that they can be used to combine features from different knowledge sources [10], as the optimum 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 $\theta_s = \{\mu_s, \Sigma_s\}$. For pooled covariance, the posterior probability becomes

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

In Equation 1, the new parameters $\lambda_s \in \mathbf{R}^\mathbf{D}$ and $\alpha_s \in \mathbf{R}$ are present in log-linear form.

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

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

for $l = 1...L_s$ mixture parameters in each class s. This is a *Hidden Conditional Random Field.*

2.1. Discriminative Training of Log-Linear Parameters

The frame level Maximum Mutual Information objective function (with regularization) is

$$\begin{aligned} \mathcal{F}_{MMI}^{(frame)}(\Lambda) &= -\tau_\Lambda ||\Lambda||^2 \\ &\quad + \sum_{r=1}^{R} \sum_{t=1}^{T_r} w_s \log p_\Lambda(s_t|x_t) \end{aligned} \tag{3}$$

$$p_\Lambda(s_t|x_t) = \frac{\exp\left(\lambda_{s_t}^\top x_t + \hat{\alpha}_{s_t}\right)}{\sum_{s'} \exp\left(\lambda_{s'}^\top x_t + \hat{\alpha}_{s'}\right)} \tag{4}$$

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 observations 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.

The sentence-level minimum phone error criterion [11] incorporates an accuracy score $A(W, W_r)$, which is the phone transcription accuracy of hypothesis sentence W given the reference sentence W_r. This is roughly equal to the number of reference phones minus the number of errors.

$$\begin{aligned} \mathcal{F}_{MPE}(\Lambda) &= -\tau_\Lambda ||\Lambda - \Lambda_0||^2 \\ &\quad + \sum_{r=1}^{R} \sum_{W \in \mathcal{M}_r} P_\Lambda(W|X_r) A(W, W_r) \end{aligned} \tag{5}$$

$$P_\Lambda(W_r|X_r) = \frac{\left(p(W_r)^{\frac{1}{\eta}} \cdot p_\Lambda(X_r|W_r)\right)^\beta}{\sum_{W \in \mathcal{M}_r} \left(p(W)^{\frac{1}{\eta}} \cdot p_\Lambda(X_r|W)\right)^\beta} \tag{6}$$

$$\begin{aligned} & p_\Lambda(X_r|W) = \\ & \max_{s_1^{T_r}|W} \left\{ \prod_{t=1}^{T_r} p(s_t|s_{t-1}) \exp\left(\lambda_{s_t}^\top x_t + \alpha_{s_t}\right) \right\} \end{aligned} \tag{7}$$

The regularization used here is called center regularization, which loosely binds Λ to their initial values Λ_0. $\mathcal{M}_r$ is the set of all possible word sequences, η is a language model scale, and β is a posterior scale.

2.2. Optimization Procedure

We use the general purpose RPROP algorithm [12] for the optimization of the objective functions in Equations 3 and 5. 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 4 the RPROP algorithm is used for optimization.

3. Discriminative Splitting

The log-linear training is only convex for a single density per state s. For mixture density training this does not hold true, and therefore the initial guess becomes very important and can dictate the final value of the objective function. 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. We adopt an approach similar to the iterative density splitting algorithm used in a maximum likelihood framework, applied to the log-linear parameters $\lambda_{s,l}$ instead of the $\mu_{s,l}$, 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. As the new lambdas iteratively adapt themselves to the training data, discriminative training of this newly split model causes an increase in the objective function .

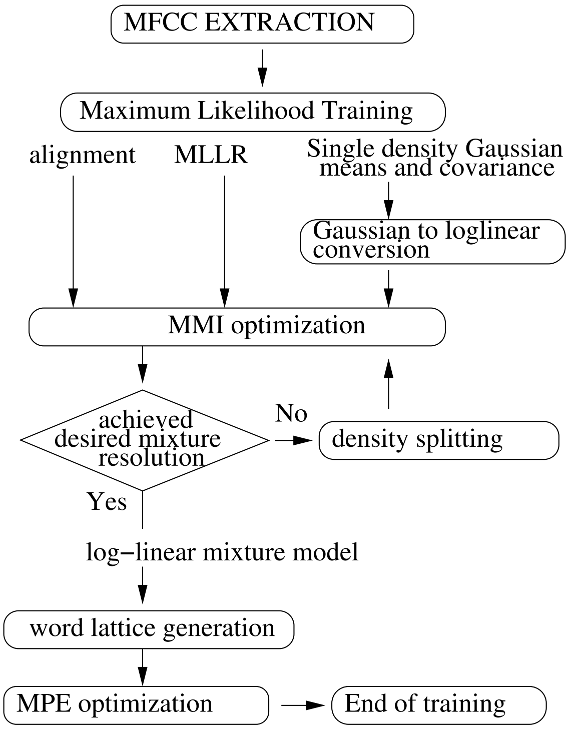

Figure 1: *Flow diagram of the combined discriminative training and splitting process*

4. Experiments and Results

4.1. Speech Corpus and Baseline System

A large vocabulary continuous speech recognition task European Parliament Plenary Sessions (EPPS) is used for discriminative training and recognition. It is a part of 2006 TC-STAR ASR evaluation campaign, and contains 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 4501 triphone CART leaves and a pooled covariance is used. The lexicon contains 54k words and there is a 4-gram language model. The initial features are 16 MFCCs (including an energy feature), augmented with a single voicedness feature. Nine such consecutive frames are concatenated together, and then projected by a classical LDA matrix to 45 dimensions. The baseline recognizer has another version with speaker adaptive training (SAT) with maximum likelihood linear regression (MLLR) and constrained MLLR (cMLLR).

4.2. Log-linear Training

A flow diagram of the training process is shown in Fig. 1. The initial alignment between the acoustic training 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 MMI 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. More details can be found in [13].

After the log-linear mixture model has reached the desired resolution, it is further trained by optimizing the MPE criterion. A language model containing only the unigrams and bigrams of the recognition LM is used for language model scores. For comparison, another system with Gaussian parameters is also trained by MPE. It is found that MPE on a Gaussian mixture model is more robust as compared to the log-linear model. This is because GMM has a constraint that the sum of weights of all the densities of a state is equal to 1. For the log-linear model this constraint is not true, and therefore we use center regularization to make it robust, as in equation 5. Another problem is the training of noise, silence and hesitation phoneme states. Log-MPE training tends to suppress them, as these states are not represented in the language model. Therefore we combine all of these noise and silence states into a single state, so that they are not differentiated from each other during the calculation of accuracy score $A(W, W_r)$.

4.3. Recognition Results

A summary of the recognition results is shown in Table 1. The WER improvement from MMI discriminative training is quite large for single densities. However, as the number of densities increases, the difference 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. 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). The real usefulness of this discriminative splitting approach is due to the improvement that it provides after a further pass of log-MPE training, which causes further reduction in WER. By this method, the total WER improvement over the baseline ML system is 1.8

Table 1: *EPPS: WER Comparison of ML split and discriminatively split mixtures.*

		WER (%)	
Discr.	**Training**	**Single**	**64 densities**
splitting	**criterion**	**density**	**per state**
Without speaker adaptive training			
No	ML	28.3	17.1
	MPE	24.5	15.8
Yes	log-MMI	23.3	16.4
	log-MPE	22.8	15.3
With SAT MLLR and cMLLR			
No	ML	16.7	13.6
	MPE	15.3	12.5
Yes	log-MMI	15.0	13.1
	log-MPE	14.7	12.1

% without SAT and 1.5 % with SAT. For comparison, we take an ML Gaussian model already split as 64 densities per state, and train it discriminatively by MPE. This is a model where only the training of means $\mu_{s,l}$ is done discriminatively, and no splitting is done in between. This way we only get a 1.3 % improvement over the ML model without SAT and 1.1 % improvement with SAT, which is less than what we obtain by integrated discriminative splitting and log-linear training . The possible reason for this could be the less susceptibility of such an approach to get stuck in a local maximum. For both Gaussian mixture model MPE and log-MPE, the same language models and lattice generation techniques are used. Also, we observe that the WER improvements with SAT are almost as good as improvements without SAT. This is because of the inclusion of MLLR matrices into the optimization feature extraction pipeline.

5. Conclusions

In this work 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. Previously it was shown that this approach provides moderate but significant gains in WER with frame-level MMI optimization. However, better criteria like MPE give better error rates with Gaussian mixture models. This paper provides some evidence to show that a combined discriminative MMI training and splitting followed by a log-linear MPE optimization gives small but consistent WER improvements in comparison to an MPE trained Gaussian mixture model. Further work is needed to increase the robustness of the log-linear lattice-based MPE training, by better regularization and special handling of noise and silence phoneme states. Also, use of better features like the state of the art MLP features needs to be investigated. Furthermore, the MLLR matrices could also be trained discriminatively to better adapt them to the rest of the system.

6. Acknowledgments

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.

7. References

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

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

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

[4] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "Mmie training of large vocabulary recognition systems," *Speech Commun.*, vol. 22, no. 4, pp. 303–314, Sep. 1997.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge, England, 2004.

[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.

[13] M. A. Tahir, R. Schlüter, and H. Ney, "Discriminative splitting of gaussian/log-linear mixture hmms for speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Hawaii, USA, Dec. 2011.