Generalized Discriminative Feature Transformation for Speech Recognition

Roger Hsiao and Tanja Schultz

InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213

wrhsiao@cs.cmu.edu, tanja@cs.cmu.edu

Abstract

We propose a new algorithm called Generalized Discriminative Feature Transformation (GDFT) for acoustic models in speech recognition. GDFT is based on Lagrange relaxation on a transformed optimization problem. We show that the existing discriminative feature transformation methods like feature space MMI/MPE (fMMI/MPE), region dependent linear transformation (RDLT), and a non-discriminative feature transformation, constrained maximum likelihood linear regression (CMLLR) are special cases of GDFT. We evaluate the performance of GDFT for Iraqi large vocabulary continuous speech recognition.

Index Terms: speech recognition, discriminative training, feature transformation

1. Introduction

Discriminative training on feature transformation has shown to be effective on improving recognition accuracy. Feature space discriminative training often involves optimization of the feature transform using some discriminative criteria such as, maximum mutual information (MMI) or minimum phone error (MPE). Well known methods include feature space MMI or MPE (fMMI/fMPE) [1][2][3] and region dependent linear transform (RDLT) [4][5]. These algorithms perform some form of linear transformation on the feature vectors and the transformation is optimized using MMI/MPE criteria.

Feature space discriminative training like fMMI/MPE or RDLT, while being effective, may come with some costs. First, the training process becomes much longer since one has to perform feature space training, followed by maximum likelihood (ML) reinitialization, and finally the model space discriminative training. Second, fMMI/MPE and RDLT relies on gradient ascent or quasi-Newton methods for optimization, which can be difficult to tune due to the complexity of the objective functions. In this paper, we propose a new optimization algorithm that is easy to tune and simplifies training process.

The proposed algorithm is named generalized discriminative feature transform (GDFT), which is inspired by our previous work on the generalized Baum-Welch (GBW) algorithm [6]. GDFT uses Lagrange relaxation to transform a well known speaker adaptation technique – constrained maximum likelihood linear regression (CMLLR) [7], which is also known as feature MLLR, into a discriminative feature transformation method. As shown in section 2, we discuss an interesting relationship between fMMI/MPE, RDLT and CMLLR. Then in section 3, we show how the formulation of GDFT can generalize these techniques. The close connection between GDFT and CMLLR also suggests a new training process which can greatly shorten the training time, and we investigate this technique in section 4. Finally, we conclude and discuss future work in section 5.

2. Review of fMMI/MPE, RDLT and CMLLR

2.1. fMMI/MPE

Feature space discriminative training algorithms like fMMI [1] or fMPE [2] perform linear transformation on feature vectors, and the transformation is often optimized for MMI/MPE objective function. The transformation is formulated as

$$z_t = x_t + Mh_t \ , \tag{1}$$

where x_t is the original feature; h_t is the Gaussian posterior vector computed by a Gaussian mixture model (GMM); M is the linear transform which is optimized for MMI/MPE objective using gradient ascent and z_t is the final feature vector. The GMM, trained from the data or induced from the acoustic model, determines the transforms applied to the feature vectors.

B. Zhang et al. in [4] shows that equation 1 can be rewritten as,

$$z_t = \sum_i \gamma_t(i)(x_t + b_i) = x_t + \sum_i \gamma_t(i) b_i \tag{2}$$

where b_i is a bias corresponding to the i-th row of M; $\gamma_t(i)$ is the posterior probability of Gaussian i at time t. From this point of view, we can consider the transformation of fMMI/MPE consists of biases only and the transformation matrix is always an identity matrix (given the mean-offset feature is not used [5]).

2.2. RDLT

RDLT [4][5] extends fMMI/MPE by allowing full transformation matrix. Hence, the final feature vectors are computed by

$$z_t = \sum_i \gamma_t(i)(A_i x_t + b_i) \ , \tag{3}$$

where A_i is the transformation matrix optimized for MMI/MPE objective. For optimization, RDLT uses a quasi-Newton algorithm which uses gradient information to approximate the Hessian matrix for performing update like Newton method.

2.3. CMLLR

CMLLR [7] is a widely used speaker adaptation algorithm. CMLLR performs linear transformation on the Gaussian means and covariances, and restricts the transformation to be the same for mean and covariance. With this restriction, [7] shows transforming the model parameters is equivalent to transforming the feature vectors as long as one subtracts $\log(|A|^2)$ to the likelihood computation, where A is the transformation matrix of the feature vectors (see equation 5).

When context expansion technique is not used or it is always fixed, the transformation matrix of fMMI/MPE is always square and identity, hence, fMMI/MPE can be considered as a model space transformation technique ($\log(|I|^2) = 0$). In contrast, RDLT is not a model space technique unless the likelihood computation is adjusted as CMLLR.

We are interested in a transformation approach similar to CMLLR, since as a model space technique, we have an option to update the transformation and the Gaussian parameters simultaneously, and it gives flexibility to the training procedure. If concurrent update of transformation parameters and Gaussian parameters is possible, it implies we can significantly reduce the training time. Also, we want the transformation optimized for an effective discriminative objective function like fMMI/MPE to improve recognition performance. In addition, we also want the transformation to be less restrictive like RDLT. Hence, we propose generalized discriminative feature transform (GDFT).

3. Generalized Discriminative Feature Transformation

In our previous work on discriminative training, we proposed GBW algorithm for discriminative training [6]. GBW uses Lagrange relaxation to optimize a transformed mutual information optimization problem. It can also be shown both Baum-Welch (BW) and extended Baum-Welch (EBW) algorithm are special cases of GBW. The same theory applies to GDFT, in which, we derive a generalized version of CMLLR which can perform ML or MMI training. The formulation is constructed in a way that GDFT can use an update equation very similar to CMLLR.

MMI optimization for GDFT, in its simplest form, can be considered as maximizing the difference between the log likelihood of the reference and the log likelihood of the competitor,

$$F(W) = Q_r(W) - Q_c(W) \,, \tag{4}$$

where W is the linear transformation of GDFT and $W \equiv [A; b]$ The subscript r and c represents reference and competitor respectively; Q is an auxiliary function to represent negative log likelihood and it is defined as

$$\begin{aligned} Q(W) &= \sum_t \sum_j \gamma_t(j)[\log(|\Sigma_j|) - \log(|A|^2) \\ &+ (W\zeta_t - \mu_j)'\Sigma_j^{-1}(W\zeta_t - \mu_j)] \,, \end{aligned} \tag{5}$$

where $\zeta_t \equiv [x_t; 1]$ is the augmented feature vector; Σ is the covariance and $\gamma_t(j)$ is the posterior probability of Gaussian j at time t. The Q function defined here is the same as the equation 57 of [7] except the terms unrelated to the optimization is removed.

It can be shown minimizing F is the same as performing MMI optimization. However, optimization of F is not trivial since the solution can be unbounded. We apply a checkpointing technique as GBW [6] which limits the changes in the scores,

$$G(W) = \sum_i |Q_i(W) - C_i| \,, \tag{6}$$

where C_i is the chosen score which we want Q_i to achieve. The function G has multiple terms since we can have multiple files in training, so we have multiple references and their corresponding competitors. As long as, the checkpoints imply higher likelihood for references and lower likelihood for competitors, minimization of G is the same as optimizing F except the limits of likelihood changes [6]. In [6], we compute Q for each word arc in the lattices, but in GDFT, we compute Q based on the whole utterance. This means for each utterance, we compute Q_r for the reference using Viterbi algorithm, Q_c for the lattice as the competitor using forward algorithm. We found that this scheme significantly improves the efficiency and does not impact performance much.

We show how to optimize equation 6. We would like to remind the readers that part of the formulation is closely related to CMLLR and readers are encouraged to read appendix C of [7] for more details. To minimize G, we first transform the problem to,

$$\begin{aligned} \min_{\epsilon, W} \quad & \textstyle\sum_i \epsilon_i \\ \text{s.t.} \quad & \epsilon_i \geq Q_i(W) - C_i \quad \forall i \\ & \epsilon_i \geq C_i - Q_i(W) \quad \forall i \,, \end{aligned}$$

where ϵ represents slack variables and i is an index to an utterance. This is equivalent to the original problem in equation 6 without constraints. We call this as the primal problem for the rest of this paper.

We can then construct the Lagrangian dual for the primal problem. The Lagrangian is defined as,

$$\begin{aligned} L^P(\epsilon, W, \alpha, \beta) &= \sum_i \epsilon_i - \sum_i \alpha_i(\epsilon_i - Q_i(W) + C_i) \\ &- \sum_i \beta_i(\epsilon_i - C_i + Q_i(W)) \end{aligned} \tag{7}$$

where $\{\alpha_i\}$ and $\{\beta_i\}$ are the Lagrange multipliers for the first and the second set of constraints of the primal problem in equation 7. The Lagrangian dual is then defined as,

$$L^D(\alpha, \beta) = \inf_{\epsilon, W} L^P(\epsilon, W, \alpha, \beta) \tag{8}$$

Now, we can differentiate L^P w.r.t. ϵ and W which includes the transformation matrix A and bias b. Hence,

$$\frac{\partial L^P}{\partial \epsilon_i} = 1 - \alpha_i - \beta_i \tag{9}$$

$$\frac{\partial L^P}{\partial W} = \sum_i (\alpha_i - \beta_i)\frac{\partial Q_i}{\partial W} \,. \tag{10}$$

By setting $\frac{\partial L^P}{\partial \epsilon_i}$ to zero, it implies,

$$\alpha_i + \beta_i = 1 \quad \forall i \,. \tag{11}$$

Assuming the covariance matrices are all diagonal, we then compute $\frac{\partial L^P}{\partial W}$ row by row,

$$\frac{\partial L^P}{\partial W_d} = \sum_i (\alpha_i - \beta_i)\frac{\partial Q_i}{\partial W_d} = -\Gamma \frac{p_d}{p_d w_d'} + w_d G^{(d)} - k^{(d)} \,, \tag{12}$$

where W_d refers to d-th row of W; $p_d = [c_{d1}, \ldots, c_{dn}, 0]$ is the extended cofactor row vector of A ($c_{ij} = cof(A_{ij})$) , and,

$$G^{(d)} = \sum_i (\alpha_i - \beta_i) \sum_j \frac{1}{\sigma_{jd}^2} \sum_t \gamma_t^i(j)\zeta_t\zeta_t' \tag{13}$$

$$k^{(d)} = \sum_i (\alpha_i - \beta_i) \sum_j \frac{1}{\sigma_{jd}^2}\mu_{jd} \sum_t \gamma_t^i(j)\zeta_t' \tag{14}$$

$$\Gamma = \sum_i (\alpha_i - \beta_i) \sum_t \sum_j \gamma_t^i(j) \,. \tag{15}$$

To solve $\frac{\partial L^P}{\partial W_d} = 0$, we can use the same method as CMLLR by first solving this quadratic equation for δ,

$$\delta^2 p_d G^{(d)-1} p_d' + \delta p_d G^{(d)-1} k^{(d)'} - \Gamma = 0 \,. \tag{16}$$

Then we can apply this update equation,

$$W_d = (\delta p_d + k^{(d)})G^{(d)-1} \ . \tag{17}$$

Updating W is an iterative process as CMLLR since the cofactors depend on other rows. As a result, we need to apply equation 17 on the whole transformation several times and recompute the cofactors until it converges. It is important to note that GDFT reduces to CMLLR if $\alpha_i = 1$ and $\beta_i = 0$ for all references and $\alpha_i = \beta_i = 0.5$ for all competitors.

Equation 12 to 17 show how W can be computed if the Lagrange multipliers, α, β, are known. In other words, W in equation 17 is a function of α and β. To estimate the multipliers, we need to construct the dual problem from the Lagrangian (equation 7), and this can be done by integrating equation 11 and 17 into equation 7. Thus, we obtain,

$$L^D(\alpha, \beta) = \sum_i (\alpha_i - \beta_i)(Q_i(W^*) - C_i) \tag{18}$$

where W^* is a function of α and β computed by equation 17. Then, we can formulate the dual problem,

$$\begin{aligned} \max_{\alpha,\beta} \quad & L^D(\alpha, \beta) \quad = \sum_i (\alpha_i - \beta_i)(Q_i(W^*) - C_i) \\ \text{s.t. } \forall i \quad & \alpha_i + \beta_i \quad = 1 \text{ and } \alpha_i, \beta_i \geq 0 \ . \end{aligned}$$

This dual problem is convex and it can be solved easily with gradient ascent. While the gradient formula can be complicated, we found that the following approximation is good enough in general,

$$\frac{\partial L^D}{\partial \alpha_i} \simeq Q_i(W^*) - C_i \ . \tag{19}$$

In theory, if the dual objective at dual optimal is smaller than the primal objective at primal optimal, there is no guarantee the dual solution is also primal optimal. Hence, using this method can only be considered as a relaxation technique, which we relax a non-convex problem into a convex one. Also, GDFT works under the EM algorithm framework, which the M-step is now replaced by solving a dual problem. To speed up the process, we can perform another EM iteration after one gradient step of equation 19. This is similar to fMMI/MPE and RDLT which we recompute the E-step after one iteration of gradient ascent or quasi-Newton method.

When GDFT is used with multiple regression classes, GDFT is the same as fMMI/MPE and RDLT which uses an GMM to compute the posterior probabilities for weighted average. However, to speed up the process, the current implementation of GDFT only uses the one transform which corresponding Gaussian yields the highest likelihood.

4. Experiments

We evaluated the performance of GDFT on a speaker dependent Iraqi ASR system with 62K vocabulary. The Iraqi system was trained with around 450 hours of audio data in force protection and medical screening domain. The acoustic model has 7000 codebooks and each codebook has at most 64 Gaussian mixtures. The model was trained with speaker adaptive training. During decoding, we performed incremental MLLR and CMLLR for adaptation. GDFT was applied on the features after adaptation. The system was evaluated in DARPA TransTac 2008 June and November evaluations as a component of the CMU English-Iraqi two-way speech-to-speech translation system [8]. In this paper, we used the June offline open set as a development set, and the November offline open set as an unseen test set. Both sets consist of conversational speech between a native English and a native Iraqi speaker and we only evaluated on the Iraqi part in this paper.

In the previous section, we learned that GDFT can be considered as a model space transformation like CMLLR. One possible way to use GDFT is to perform feature transformation and model space discriminative training at the same time. This form of training is similar to speaker adaptive training using CMLLR, except we perform both transformation and model updates at the same time. We call this form of training as joint training in this paper. In the following experiments, we chose boosted MMI (bMMI) [1] for the model space discriminative training.

In the first experiment, we evaluated the performance of GDFT in the form of joint training and compare to using only the model space training. For GDFT, we used 16 transforms and they are optimized for MMI objective. The checkpoints are set to be 10% higher for reference and 10% lower for competitor in terms of log likelihood. During each iteration, we collected statistics for GDFT and bMMI for joint training and updated the acoustic model and the feature transform concurrently.

Iter	ML	1	2	3	4
bMMI	37.0	36.1	35.9	35.6	**35.4**
GDFT+bMMI	37.0	35.8	35.0	**34.4**	35.0

Table 1: WER(%) of bMMI and GDFT+bMMI joint training on dev set. Each iteration costs almost the same amount of time for both methods.

Table 1 shows the joint training using GDFT and bMMI can improve training using bMMI only. The difference in performance is one point absolute. This indicates joint training using GDFT and bMMI is effective and has the potential to shorten the time for discriminative training, since it is no longer necessary to separate feature space and model space training, and the additional computation for GDFT is the same as CMLLR which is neglible. For the next experiment, we examined different training procedures which includes the joint training (GDFT+bMMI), the conventional feature space training followed by model space training (GDFT $\rightarrow$ bMMI), and model space training followed by feature space training (bMMI $\rightarrow$ GDFT). Figure 1 shows that the performance of different training procedures. We found that joint training (GDFT+bMMI)

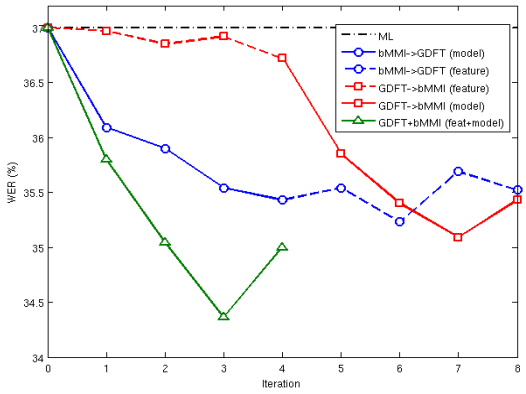

Figure 1: WER of different training procedures on dev set. Each iteration costs almost the same amount of time for all procs.

outperformed other training procedures and it achieved its best performance earlier than the other methods as well. GDFT → bMMI gives 35.1% WER which is better than using bMMI alone (35.4%). It is interesting to see although using GDFT alone gave little improvement, the output features helped bMMI to improve the overall performance in the later stages of training. bMMI → GDFT is unstable as shown in the figure although we observed continual improvement on the objective function. This probably implies overfitting. Figure 2 shows performance

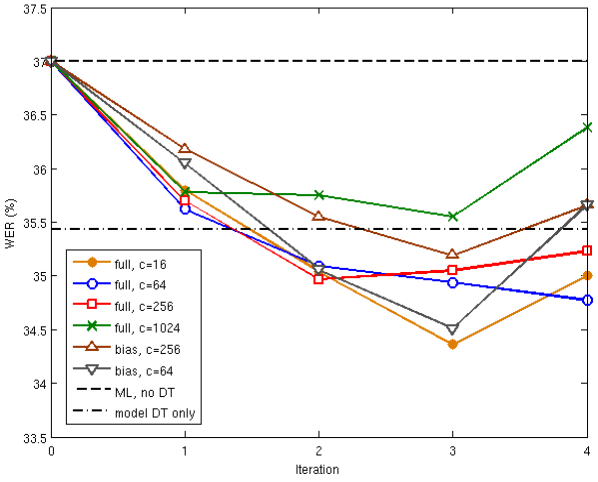

Figure 2: Performance of GDFT and bMMI joint training using different configurations.

of joint training using different configurations. In the figure, "full" means GDFT used full transformation and "bias" means GDFT used bias only. Different number of regression classes was tested. In the experiment, we found that GDFT does not need many classes for optimal performance. When only bias is used for feature transform, GDFT is the same as fMMI except the optimization algorithm is different, and in the current implementation of GDFT, it only selects one transform based on the best likelihood scores, but not a weighted average using posterior as the original fMMI/MPE. This result is different from what we expected since it is generally believed fMMI/MPE needs more than a thousand transforms [3]. More experiments need to be done in order to identify whether the small number of transforms preferred by GDFT is due to the optimization algorithm, the training procedure or how the transforms are selected. Nonetheless, if GDFT can reduce the amount of transforms required, it is beneficial since it saves the computation. In sum, GDFT with 16 full transforms or GDFT with 64 bias only transforms give 34.4% and 34.5% WER respectively, which is better than 35.4% WER by using bMMI model space training only.

Finally, we compared the performance on the unseen test set, which is the TransTac 2008 November open set. The result is in table 2 and we observed GDFT+bMMI joint training with 16 full transforms gave the best performance.

5. Conclusion and Future Work

In this paper, we introduce GDFT which uses Lagrange relaxation to construct a discriminative version of CMLLR. GDFT is a generalization of CMLLR and fMMI/MPE and very similar to the RDLT. With joint training with bMMI model space discriminative training, GDFT can improve the ML and bMMI

	WER (%)	Rel. imprv.
ML	35.7	-
bMMI	34.3	3.9%
GDFT+bMMI (bias, c=64)	33.7	5.6%
GDFT+bMMI (full, c=16)	33.2	7.0%

Table 2: WER(%) of bMMI and GDFT+bMMI joint training on the unseen TransTac Nov08 open set.

baseline. The results also suggest that GDFT can shorten the training time, since it is no longer necessary to separate the feature space and model space training to exploit the benefits of discriminative feature transformation. For the future work, more experiments will be done to compare different optimization algorithms used for feature space discriminative training, and we will study possible smoothing techniques for GDFT.

6. Acknowledgments

This work is in part supported by US DARPA under the TransTac (Spoken Language Communication and Translation System for Tactical Use) program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

7. References

[1] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for Model and Feature-space Discriminative Training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[2] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and Geoffrey Zweig, "fMPE: Discriminatively Trained Features for Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[3] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proceedings of the INTERSPEECH*, 2005.

[4] B. Zhang, S. Matsoukas, and R. Schwartz, " Discriminatively Trained Region Dependent Feature Transforms For Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[5] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent Progress on the Discriminative Region-dependent Transform for Speech Feature Extraction," in *Proceedings of the INTERSPEECH*, 2006.

[6] R. Hsiao, Y.C. Tam, and T. Schultz, "Generalized Baum-Welch Algorithm for Discriminative Training on Large Vocabulary Continuous Speech Recognition System," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[8] N. Bach, M. Eck, P. Charoenpornsawat, T. Köhler, S. Stüker, T. Nguyen, R. Hsiao, A. Waibel, S. Vogel, T. Schultz, and A. W. Black, "The CMU TransTac 2007 Eyes-free, and Hands-free Two-way Speech-to-speech Translation System," in *Proceedings of the IWSLT*, 2007.