Improvements to Generalized Discriminative Feature Transformation for Speech Recognition

Roger Hsiao, Florian Metze and Tanja Schultz

InterACT, Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
{wrhsiao, fmetze, tanja}@cs.cmu.edu

Abstract

Generalized Discriminative Feature Transformation (GDFT) is a feature space discriminative training algorithm for automatic speech recognition (ASR). GDFT uses Lagrange relaxation to transform the constrained maximum likelihood linear regression (CMLLR) algorithm for feature space discriminative training. This paper presents recent improvements on GDFT, which are achieved by regularization to the optimization problem. The resulting algorithm is called regularized GDFT (rGDFT) and we show that many regularization and smoothing techniques developed for model space discriminative training are also applicable to feature space training. We evaluated rGDFT on a real-time Iraqi ASR system and also on a large scale Arabic ASR task.
Index Terms: speech recognition, discriminative training

1. Introduction

We introduced generalized discriminative feature transformation (GDFT) for feature space discriminative training in [1]. GDFT uses Lagrange relaxation to transform the constrained maximum likelihood linear regression (CMLLR) algorithm to perform feature space discriminative training. In our previous work, we showed that GDFT transforms the features in a way similar to feature space MMI/MPE (fMMI/MPE) [2] and region dependent linear transformation (RDLT) [3], and yet, GDFT allows joint optimization of model space and feature space which can cut the overall training time by half.

In this paper, we present recent improvements on GDFT, where we incorporate regularization to the GDFT algorithm. The resulting algorithm is called regularized GDFT (rGDFT). Adding regularization to feature space discriminative training is not new. In [4], regularization is applied to the large margin based semi-tied covariance (STC) transforms, which is equivalent to some form of RDLT. However, the advantage of rGDFT is that it requires little modification to the CMLLR update equations, and its regularization is based on a simple distance measure which can be extended easily. In this paper, we show that rGDFT can outperform GDFT in various recognition tasks, and we discuss how existing smoothing techniques for model space discriminative training can be applied to feature space training.

This paper is organized as follows: In section 2, we review GDFT and describe how regularization can be incorporated. Section 3 contains the experiments for comparing rGDFT and GDFT and discusses the differences of these feature space discriminative training algorithms. We then show the performance of rGDFT on a large scale Arabic speech recognition system in section 4. Finally, we conclude and discuss future work in section 5.

2. Generalized Discriminative Feature Transformation with Regularization

This section reviews GDFT and explains how we can add regularization to GDFT. To simplify the discussion, we start from the MMI objective function, but it is easy to generalize it for other discriminative objective functions. 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,

$$F(W) = Q_r(W) - Q_c(W)\,, \tag{1}$$

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

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

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 equivalent to the auxiliary function used in CMLLR [5] except the terms unrelated to the optimization are removed.

Optimizing F is not trivial since the solution can be unbounded for some parameters. However, this issue can be handled by converting the problem into,

$$G(W) = \sum_i |Q_i(W) - C_i|\,, \tag{3}$$

where C_i is the target score that 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 competing hypotheses. As long as the target values 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. While we can compute Q for the whole utterance or each word arc in the lattices, we chose the utterance level since it is more efficient [1].

Equation 3 is the objective function for the original GDFT. One can easily extend GDFT by incorporating a regularization term to G in equation 3. The objective function of rGDFT is

$$G'(W) = \sum_i |Q_i(W) - C_i| + \frac{D}{2}||W - W^0||_F^2\,, \tag{4}$$

where $W^0 \equiv [A^0; b^0]$ is the backoff transform that we want W to backoff; $||W - W^0||_F$ is the Frobenius norm between W and W^0 and D is a tunable parameter controlling the weight of this regularization term. When $D = 0$, rGDFT reduces to GDFT.

We show how to optimize equation 4. The formulas are closely related to CMLLR. Details of CMLLR formulation is available in the appendix C of [5]. To minimize G', we first transform the problem to,

$$\begin{array}{ll} \min\limits_{\epsilon,W} & \sum_i \epsilon_i + \frac{D}{2}||W - W^0||_F^2 \\ \text{s.t.} & \epsilon_i \geq Q_i(W) - C_i \qquad \forall i \\ & \epsilon_i \geq C_i - Q_i(W) \qquad \forall i \ , \end{array}$$

where ϵ represents slack variables and i is an index to an utterance. This is equivalent to the original problem in equation 4 without constraints. We call this as the primal problem for the rest of this paper.

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

$$\begin{aligned} L^P(\epsilon, W, \alpha, \beta) &= \sum_i \epsilon_i - \sum_i \alpha_i(\epsilon_i - Q_i(W) + C_i) \\ &- \sum_i \beta_i(\epsilon_i - C_i + Q_i(W)) \\ &+ \frac{D}{2}\sum_d (w_d - w_d^0)(w_d - w_d^0)' \end{aligned} \quad (5)$$

where w_d and w_d^0 represent the d-th row of W and W^0 respectively; $\{\alpha_i\}$ and $\{\beta_i\}$ are the Lagrange multipliers for the first and the second set of constraints of the primal problem in equation 5. The Lagrangian dual is then defined as,

$$L^D(\alpha, \beta) = \inf_{\epsilon, W} L^P(\epsilon, W, \alpha, \beta) \quad (6)$$

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

$$\frac{\partial L^P}{\partial \epsilon_i} = 1 - \alpha_i - \beta_i \quad (7)$$

$$\frac{\partial L^P}{\partial W} = \sum_i (\alpha_i - \beta_i)\frac{\partial Q_i}{\partial W} + \frac{D}{2}\frac{\partial ||W - W^0||_F^2}{\partial W} \ . \quad (8)$$

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

$$\alpha_i + \beta_i = 1 \quad \forall i \ . \quad (9)$$

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

$$\begin{aligned} \frac{\partial L^P}{\partial w_d} &= \sum_i (\alpha_i - \beta_i)\frac{\partial Q_i}{\partial w_d} + D(w_d - w_d^0) \\ &= -\Gamma \frac{p_d}{p_d w_d'} + w_d(G^{(d)} + DI) - (k^{(d)} + D w_d^0) \end{aligned} \quad (10)$$

where $p_d = [c_{d1}, \ldots, c_{dn}, 0]$ is the extended cofactor row vector of A ($c_{ij} = cof(A_{ij})$) , and,

$$G^{(d)} = \sum_i (\alpha_i - \beta_i) \sum_j \frac{1}{\sigma_{jd}^2} \sum_t \gamma_t^i(j)\zeta_t\zeta_t' + DI \quad (11)$$

$$k^{(d)} = \sum_i (\alpha_i - \beta_i) \sum_j \frac{\mu_{jd}}{\sigma_{jd}^2} \sum_t \gamma_t^i(j)\zeta_t' + D w_d^0 \quad (12)$$

$$\Gamma = \sum_i (\alpha_i - \beta_i) \sum_t \sum_j \gamma_t^i(j) \ . \quad (13)$$

The only difference between GDFT and rGDFT is the definition of G and k. In rGDFT, some additional counts are added to the accumulators. This additional computation is negligible. In fact, by adding $D \times I$ to G, as long as D is large enough, it helps G to have enough rank for inversion and this also reduces possible numerical issues.

There are many possible choices of W^0. The simplest case is the identity matrix, I, which is the one we used in our experiments. Other possible choices include ML estimates, MMI estimates or the transform from the previous iteration. The D parameter serves as the same purpose as the D-term used in [4] and we apply the same heuristics, i.e. $D = E \times \gamma_{den}$. When there are more than one regression classes, we have one D value for each transform, which is like one D value for each Gaussian in EBW or GBW.

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,

$$\delta^2 p_d G^{(d)-1} p_d' + \delta p_d G^{(d)-1} k^{(d)'} - \Gamma = 0 \ . \quad (14)$$

Then we can apply this update equation,

$$w_d = (\delta p_d + k^{(d)}) G^{(d)-1} \ . \quad (15)$$

Updating W is an iterative process as CMLLR since the cofactors depend on other rows. As a result, we need to apply equation 15 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 10 to 15 show how W can be computed if the Lagrange multipliers, α, β, are known. In other words, W in equation 15 is a function of α and β. To estimate the multipliers, we need to construct the dual problem from the Lagrangian (equation 5), and this can be done by integrating equation 9 and 15 into equation 5. Thus, we obtain,

$$L^D(\alpha, \beta) = \sum_i (\alpha_i - \beta_i)(Q_i(W^*) - C_i) \quad (16)$$

where W^* is a function of α and β computed by equation 15. Then, we can formulate the dual problem,

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

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,

$$\frac{\partial L^D}{\partial \alpha_i} \simeq Q_i(W^*) - C_i \ . \quad (17)$$

As in [1], rGDFT does not fulfill the strong duality condition, so using this method can only be considered as a relaxation technique, which we relax a non-convex problem into a convex one. The training procedure of rGDFT also remains the same as GDFT as described in [1].

When GDFT is used with multiple regression classes, GDFT is the same as fMMI/MPE and RDLT which uses a Gaussian mixture model (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. By doing so, GDFT is equivalent to a model space transformation technique which selects a transform based on the incoming feature vectors.

3. Experiments on GDFT and rGDFT

We evaluated the performance of GDFT and rGDFT 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 system has 300K Gaussian distributions and it is optimized for real-time performance with aggressive pruning. Incremental CMLLR and MLLR are used for speaker adaptation. Detailed system description is available in [1, 6] and same as the previous experiments, we used the Jun08 open set as a development set, and the Nov08 open set as an unseen test set. Both sets consist of conversational speech between a native English and Iraqi speaker and we only evaluated on the Iraqi part in this paper.

As in [1], we are interested in the training procedure of GDFT and we want to investigate how GDFT can be incorporated with the model space discriminative training. For model space training, we chose boosted MMI (BMMI) [7] which is also the objective function used for GDFT. For the training procedure, we explored the joint training, (r)GDFT+BMMI, which optimizes the feature transforms and the model parameters simultaneously, and we also tried the conventional approach which we optimize for the feature space first, then use the new features for model space training. This conventional procedure is denoted as (r)GDFT→BMMI.

Table 1 is the comparison of GDFT and rGDFT using different training procedures. The training in this experiment only consists of feature space training. The acoustic model is the ML model. For rGDFT, the regularization parameter E is set to one. From the results, we observed that regularization allows GDFT to use more transforms. The performance of GDFT degrades when there are 1024 transforms and the training fails for 2048 transforms. However, rGDFT continues to improve the ML baseline with more than 1024 transforms while it has the same performance when there are only 16 transforms. In this experiment, rGDFT achieved 35.8% WER with 2048 transforms, which is better than the ML baseline with 1.2% absolute improvement.

Training proc.	# transforms	WER (%)
ML	-	37.0
GDFT	16	36.7
GDFT	1024	38.5
GDFT	2048	-
rGDFT	16	36.7
rGDFT	1024	36.1
rGDFT	2048	**35.8**
rGDFT	4096	35.9

Table 1: WER(%) of GDFT and rGDFT(E=1) on the dev set (TransTac Jun08 open set).

Then, we explored how the regularization parameter, E, may affect the performance of rGDFT. We tested $E = 0$ which means no regularization, and E from one to two. When there are 1024 transforms, GDFT, that is rGDFT without regularization ($E = 0$), degrades the performance and has a WER of 38.5%. rGDFT, however, has the same WER of 36.1% for different E from one to two. Similarly, when there are 2048 transforms, GDFT degrades the performance, but rGDFT can outperform the baseline system with a WER of 35.8% for E from one to two. The performance of rGDFT is not sensitive to the choice of E which means tuning should be easy.

Then, we tested rGDFT with model space discriminative training. Similar to the experiments we conducted, we evaluated different training procedures. In this experiment, E is set to be one for rGDFT. Figure 1 shows that the joint rGDFT and BMMI training can outperform GDFT and BMMI joint training. Hence, regularization proves to be helpful for GDFT. The rGDFT → BMMI training achieved 33.8% WER which is the best number on this development set among all settings. However, joint training, rGDFT+BMMI, has a comparable WER of 34.0% while the training time is only half of the rGDFT → BMMI training. Basically, each procedure spends roughly the same amount of time for each iteration, and each iteration needs about 30 hours to process the whole train set on 10 machines.

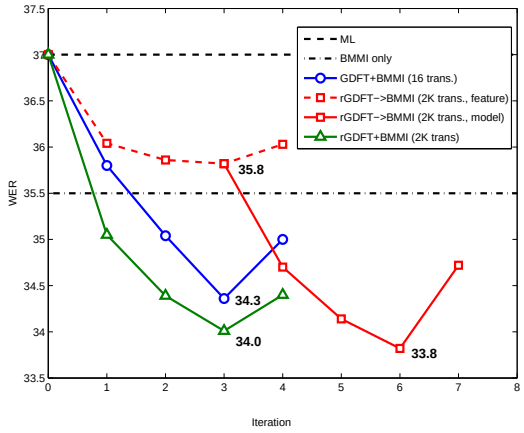

Figure 1: WER of different training procedures using rGDFT on dev set (TransTac Jun08 open set).

Finally, table 2 compares the performance of GDFT and rGDFT on the unseen TransTac Nov08 open test set. The models are tuned on the dev set and tested on this unseen test set. The results conclude that rGDFT when combined with BMMI is better than GDFT with BMMI joint training. Unlike the results from the dev set, rGDFT+BMMI joint training is better than the rGDFT → BMMI training. The reason is that the performance of rGDFT+BMMI joint training at different grammar factor and word penalty settings has a smaller variance compared to the rGDFT → BMMI training on the dev set..

	# trans.	WER (%)	Rel. imprv.
ML	-	35.7	-
BMMI	-	34.3	3.9%
GDFT+BMMI	16	33.2	7.0%
rGDFT+BMMI	2048	32.0	10.4%
rGDFT → BMMI	2048	32.6	8.7%

Table 2: Performance of GDFT and rGDFT on the unseen TransTac Nov08 open set.

4. Experiments on Large System

This section describes the performance of GDFT on the CMU speech recognition system for Modern Standard Arabic (MSA) [8], as it was developed for the GALE 2009 Speech-to-Text

evaluation. Unlike the Iraqi ASR system in section 3, this system is optimized for recognition performance without the real-time constraint.

This Arabic system is trained on approximately 1150 hours of training data, taken from the GALE P2 and P3 sets using both a vowelized, and an unvowelized dictionary. The training data provides manual segmentation and speaker clusters, while for the testing data, clusters have been generated automatically. We also use Bottleneck features [9] in order to build diverse systems for cross-adaptation and combination.

The feature extraction process is described in [8]. It computes 13 Mel-Frequency Cepstral Coefficients (MFCC) per frame. Then we concatenated 15 adjacent MFCC frames and performed Linear Discriminate Analysis (LDA) to project the 195 dimensional feature vectors into a 42 dimensional space. For Bottleneck based systems, the LDA transform is replaced by the 3 layer feed-forward part of the multi-layer perceptron (MLP) using a 195-3000-42 architecture, followed by stacking of 9 consecutive Bottleneck frames. A 42-dimensional feature vector is again generated by LDA, followed by STC.

For the development of our GMM based context dependent acoustic models, we applied an entropy-based polyphone decision tree to cluster the quinphones with context width ± 2. The system uses 6000 phonetically tied quinphones with at most 150 Gaussians per state, assigned using merge and split training, with diagonal covariance matrices.

The language model (LM) is trained from a variety of sources. The Arabic Gigaword corpus distributed by LDC is the major text resource for language modeling. In addition, we harvested transcripts from Al-Jazeera, Al-Akhbar, and Akhbar Elyom, as described in [8]. Acoustic transcripts from FBIS, TDT-4, GALE broadcast news (BN) and broadcast conversations (BC) were also used. The total number of words in the corpus amounted to 1.1 billion. The final LM is a 4-gram LM with 692M n-grams and 737K words in vocabulary.

The system uses three passes: 1) unvowelized speaker independent (UnVow SI) decoding using a two-stream MFCC+MLP system, 2) unvowelized speaker adaptive (UnVow SA) decoding using an MFCC system and the UnVow SI hypotheses for adaptation, and 3) vowelized speaker adaptive (Vow SA) decoding using the UnVow SA hypotheses for adaptation.

In our experiments, discriminative training was performed on the UnVow SA and the Vow SA systems. The GALE dev07 and dev08 sets were used as development sets and the eval08 set was used as an unseen test set. Table 3 shows the performance of discriminative training on the UnVow SA and the Vow SA system. For rGDFT, 2048 transforms were used and E is set to one. The improvement of WER for rGDFT+BMMI on the UnVow SA system is comparable to the improvement we observed in the Iraqi ASR system. While the improvement on the Vow SA system is small, more investigation is needed in the future to understand what contributes to the smaller improvement on this system.

5. Conclusion and Future Work

In this paper, we introduce rGDFT which improves GDFT for feature space discriminative training. The formulation of rGDFT shows the regularization of feature space training can be done in various ways, and techniques developed for model space discriminative training also applicable to rGDFT. For future work, we will explore different backoff schemes for rGDFT as well as I-smoothing.

	system	dev07	dev08	eval08
ML	UnVow SA	16.7	19.3	16.1
BMMI	UnVow SA	15.7	18.1	15.5
rGDFT+BMMI	UnVow SA	15.0	17.7	15.2
ML	Vow SA	14.3	15.9	13.9
rGDFT+BMMI	Vow SA	13.7	15.3	13.3

Table 3: WER(%) of the UnVow SA and Vow SA systems on the GALE test sets.

6. Acknowledgments

This work is in part supported by US DARPA under the TransTac (Spoken Language Communication and Translation System for Tactical Use) program, and the GALE (Global Autonomous Language Exploitation) program under Contract No. HR0011-06-2-0001. 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] R. Hsiao and T. Schultz, "Generalized Discriminative Feature Transformation for Speech Recognition," in *Proceedings of the INTERSPEECH*, 2009, pp. 664–667.

[2] D. Povey, "Improvements to fMPE for Discriminative Training of Features," in *Proceedings of the INTERSPEECH*, 2005, pp. 2977–2980.

[3] 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, pp. 1573–1576.

[4] G. Saon, D. Povey, and H. Soltau, "Large Margin Semi-tied Covariance Transforms for Discriminative Training," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 3753–3756.

[5] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[6] 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.

[7] 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, pp. 4057–4060.

[8] M. Noamany, T. Schaaf, and T. Schultz, "Advances in the CMU/InterACT Arabic GALE Transcription System," in *Proceedings of HLT/NAACL*, 2007, pp. 129–132.

[9] F. Grézl and P. Fousek, "Optimizing Bottle-neck Features for LVCSR," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 4729–4732.