DISCRIMINATIVE FEATURE TRANSFORMATION BY GUIDED DISCRIMINATIVE TRAINING

Roger Hsiao and *Brian Mak*

Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{hsiao,mak}@cs.ust.hk

ABSTRACT

In this paper, we investigate *guided discriminative training* in the context of improving multi-class classification problems. We are interested in applications that require improvement in the classification performance of only a subset of the classes at the possible expense of poorer classification performance of the remaining classes. However, should the classification of the remaining classes deteriorate, it is guaranteed not to be worse than the extent that the user specifies. The problem is formulated as a nonlinear programming problem, which can be translated to a unconstrained nonlinear optimization problem using the barrier method that, in turn, can be solved by gradient descent method. To prove the concept, we apply guided discriminative training to derive an optimal linear transformation on the mel-filterbank log power spectra to improve TIMIT phoneme classification. Encouraging results are obtained.

1. INTRODUCTION

Many problems related to multi-class classification or recognition, such as data modeling and model adaptation, are cast as optimization problems. Common approaches to solve these optimization problems include the maximum likelihood (ML), the maximum a posteriori (MAP), and the minimum classification errors (MCE) methods. Usually these methods are applied to optimize *all* the different classes simultaneously. However, there are situations where the classification performance of only a subset of the classes needs improvement. For example, after building the acoustic models of various phonemes in a computer-aided language learning (CALL) application, some phoneme pairs, such as "f" and "v", or "sh" and "zh", are highly confusable. Thus, it would be good to boost the discrimination between these pairs of phonemes. Although some sort of discriminative training may be applied (e.g. corrective training [1]) to remedy the situation, in general, it is hard to improve the discrimination between only a subset of classes without sacrificing the discrimination with the remaining classes.

In this paper, we investigate a special form of discriminative training which we call *guided discriminative training* (GDT) in the context of multi-class classification problems. We are interested in the following problem: Can we improve the classification performance of only a subset of the classes, which we call the "discriminatory group", at the expense of *possibly*, (but not *necessarily*), a poorer classification performance of the remaining classes, which we call the "constrained group"? Also, should the classification of the constrained group get worse, can we guarantee that it will not be worse than to the extent that a user may specify? Our work is inspired by Xing's work on distance metric learning [2] in which a distance metric is learned from data to improve subsequent clustering on the data using side-information provided by the user. Here, the user has to provide three pieces of side-information to our GDT method:

- membership of the discriminatory group
- membership of the constrained group
- the maximum allowable performance degradation of the constrained group.

Our GDT method is also similar to corrective training except that the latter does not guarantee the performance of the constrained group.

To prove the concept, we apply our guided discriminative training to derive an optimal linear transformation on the mel-filterbank log power spectra to improve TIMIT phoneme classification. Traditionally, statistical techniques such as linear discriminant analysis (LDA) [3] and heteroscedastic discriminant analysis (HDA) [4] are applied to achieve feature transformation. More recently, MCE discriminative training [5] has been used to optimize a feature transformation on log power spectra [6], mel-filterbank log power spectra [7], or LPCC [8] with the final goal of improving speech recognition accuracy. On the other hand, discriminative feature extraction [9, 10, 11] has also been proposed to derive discriminative features through modifying the auditory-based filters in the filterbank-based feature extraction process.

The paper is organized as follows. In the next section, we formulate the feature transformation problem in our guided discriminative training framework as a nonlinear programming problem. In Section 2.2, we show how this can be solved by the barrier method. This is followed by the experimental evaluation in Section 3 and conclusions in Section 4.

2. DISCRIMINATIVE FEATURE TRANSFORMATION

Similar to the work in [7], we would like to generalize the discrete cosine transform (DCT) commonly used to generate mel-

frequency cepstral coefficients (MFCCs) by a linear transformation optimally derived in a data-driven approach. However, while [7] computed the optimal linear transformation using MCE training, we derive it using our guided discriminative training (GDT) method.

Let us denote the mel-filterbank log power spectral vector by $\mathbf{y} \in \mathbb{R}^N$, the generalized linear transformation by an $M \times N$ matrix $\mathbf{A}^T$ ($M \leq N$), and the MFCC vector by $\mathbf{x} \in \mathbb{R}^M$. Thus, we have $\mathbf{x} = \mathbf{A}^T\mathbf{y}$ for the generalized MFCCs. Also let ω_i and ω_j be two acoustic classes in a multi-class classification problem in speech, and let $E_{\mathbf{A}}(i,j)$ be an estimate of the classification error between ω_i and ω_j after feature transformation using $\mathbf{A}$. All class pairs are divided into three groups:

- the discriminatory group $\mathcal{D}$: $(\omega_i, \omega_j) \in \mathcal{D}$ if ω_i and ω_j are to be more discriminative with each other.
- the constrained group $\mathcal{C}$: $(\omega_i, \omega_j) \in \mathcal{C}$ if the discrimination between ω_i and ω_j may be sacrificed.
- the don't-care group $\mathcal{R}$: $(\omega_i, \omega_j) \in \mathcal{R}$ if we do not care about the discrimination between ω_i and ω_j after the feature transformation.

The problem of feature transformation may then be formulated under the GDT framework as follows:

$$\begin{aligned} &\text{Minimize} \quad \sum_{(\omega_i,\omega_j)\in\mathcal{D}} E_{\mathbf{A}}(i,j) \\ &\text{subject to} \quad \frac{E_{\mathbf{W}}(k,l)}{E_{\mathbf{A}}(k,l)} \geq \delta \ , \ \forall(\omega_k,\omega_l) \in \mathcal{C} \ , \end{aligned} \tag{1}$$

where $\mathbf{W}^T$ is the DCT transform with

$$W^T_{mn} = \sqrt{\frac{2}{N}} \cos\left[(n-0.5)\frac{m\pi}{N}\right] \ ,$$

$1 \leq m \leq M$ and $1 \leq n \leq N$, and $\delta \in [0,1)$ will be called the *degradation level* of the constrained group.

The meaning of the above expression is that the GDT method tries to find a linear feature transformation $\mathbf{A}$ which will minimize an estimate of the classification errors between classes in the discriminatory group, and at the same time, will not increase an estimate of the classification error between any two classes in the constrained group by more than $\frac{100}{\delta}\%$ of the error given by the DCT transform.

2.1. Estimate of the Classification Error

If we assume that all classes are equally likely and are Gaussian distributed so that $P(\mathbf{x}^{(k)}) \sim N(\mathbf{x}; \boldsymbol{\mu}_x^{(k)}, \boldsymbol{C}_x^{(k)}), \forall\omega_k$, then we may use the Bhattacharyya bound

$$E_{\mathbf{A}}(i,j) = 0.5e^{-d_{\mathbf{A}}(i,j)} \tag{2}$$

as an estimate for the classification error between ω_i and ω_j. $d_{\mathbf{A}}(i,j)$ is the Bhattacharyya distance between the two classes, and is given by

$$\begin{aligned} d_{\mathbf{A}}(i,j) &= \frac{1}{8}\underbrace{(\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)})^T \mathbf{B}_x^{(ij)^{-1}} (\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)})}_{f(\mathbf{A})} \\ &\quad + \frac{1}{2}\underbrace{\ln \frac{\left|\mathbf{B}_x^{(ij)}\right|}{\sqrt{|\boldsymbol{C}_x^{(i)}||\boldsymbol{C}_x^{(j)}|}}}_{g(\mathbf{A})} \end{aligned} \tag{3}$$

where $\mathbf{B}_x^{(ij)} = \frac{\boldsymbol{C}_x^{(i)} + \boldsymbol{C}_x^{(j)}}{2}$. (Note that a subscript of x or y is used to denote a quantity related to the corresponding random variable in the cepstral domain or in the filterbank log power spectral domain respectively.)

2.2. Barrier Method

The feature transformation problem in Eqn (1) is a nonlinear programming problem, which can be solved by the barrier method. The barrier method converts the constrained optimization problem of Eqn (1) into the following unconstrained optimization problem: Minimize

$$\begin{aligned} F(\mathbf{A},\beta) &= 0.5 \sum_{(\omega_i,\omega_j)\in\mathcal{D}} e^{-d_{\mathbf{A}}(i,j)} \\ &\quad - \beta \sum_{(\omega_k,\omega_l)\in\mathcal{C}} \log\left(\frac{e^{-d_{\mathbf{W}}(k,l)}}{e^{-d_{\mathbf{A}}(k,l)}} - \delta\right) \ . \end{aligned} \tag{4}$$

2.2.1. Derivatives for Gradient Descent Method

The optimal linear transformation $\mathbf{A}$ in the unconstrained function $F(\mathbf{A},\beta)$ of Eqn (4) may now be solved by the iterative gradient descent method. Differentiating $F(\mathbf{A},\beta)$ w.r.t. $\mathbf{A}$, we obtain

$$\begin{aligned} \frac{\partial F(\mathbf{A},\beta)}{\partial \mathbf{A}} &= -0.5 \sum_{(\omega_i,\omega_j)\in\mathcal{D}} \frac{\partial d_{\mathbf{A}}(i,j)}{\partial \mathbf{A}} e^{-d_{\mathbf{A}}(i,j)} \\ &\quad - \beta \sum_{(\omega_k,\omega_l)\in\mathcal{C}} \frac{\partial d_{\mathbf{A}}(l,k)}{\partial \mathbf{A}} \frac{e^{-d_{\mathbf{W}}(k,l)}}{e^{-d_{\mathbf{W}}(k,l)} - \delta e^{-d_{\mathbf{A}}(k,l)}} \end{aligned} \tag{5}$$

From Eqn (3), we get

$$\frac{\partial d_{\mathbf{A}}}{\partial \mathbf{A}} = \frac{1}{8} \cdot \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}} + \frac{1}{2} \cdot \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}} \ . \tag{6}$$

Since $\mathbf{x} = \mathbf{A}^T\mathbf{y}$, we have, for any class ω_k

$$\boldsymbol{\mu}_x^{(k)} = \mathbf{A}^T\boldsymbol{\mu}_y^{(k)} \quad \text{and} \quad \boldsymbol{C}_x^{(k)} = \mathbf{A}^T\boldsymbol{C}_y^{(k)}\mathbf{A} \ . \tag{7}$$

Taking the derivatives of $f(\mathbf{A})$ and $g(\mathbf{A})$ in Eqn (3) w.r.t. $\mathbf{A}$, and making use of the relations in Eqn (7), we get

$$\begin{aligned} \frac{\partial g(\mathbf{A})}{\partial \mathbf{A}} &= 2\mathbf{B}_y^{(ij)}\mathbf{A}\mathbf{B}_x^{(ij)^{-1}} \\ &\quad - \left(\boldsymbol{C}_y^{(i)}\mathbf{A}\boldsymbol{C}_x^{(i)^{-1}} + \boldsymbol{C}_y^{(j)}\mathbf{A}\boldsymbol{C}_x^{(j)^{-1}}\right) \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial f(\mathbf{A})}{\partial \mathbf{A}_{mn}} &= 2(\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)})^T \mathbf{B}_x^{(ij)^{-1}} \frac{\partial(\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)})}{\partial \mathbf{A}_{mn}} \\ &\quad + (\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)})^T \frac{\partial \mathbf{B}_x^{(ij)^{-1}}}{\partial \mathbf{A}_{mn}} (\boldsymbol{\mu}_x^{(i)} - \boldsymbol{\mu}_x^{(j)}) \end{aligned} \tag{9}$$

where, if $\boldsymbol{u}_n$ represents the nth basis (unit) vector, then

$$\frac{\partial \boldsymbol{\mu}_x^{(i)}}{\partial \mathbf{A}_{mn}} = \mu_{y,m}^{(i)} \boldsymbol{u}_n \tag{10}$$

$$\frac{\partial \mathbf{B}_x^{(ij)^{-1}}}{\partial \mathbf{A}_{mn}} = -\mathbf{B}_x^{(ij)^{-1}} \frac{\partial \mathbf{B}_x^{(ij)}}{\partial \mathbf{A}_{mn}} \mathbf{B}_x^{(ij)^{-1}} \tag{11}$$

$$\frac{\partial (\mathbf{B}_x^{(ij)})_{pq}}{\partial \mathbf{A}_{mn}} = \begin{cases} 0 & \text{if } p \neq n, q \neq n \\ (\mathbf{B}_y^{(ij)}\mathbf{A})_{mq} & \text{if } p = n, q \neq n \\ (\mathbf{B}_y^{(ij)}\mathbf{A})_{mp} & \text{if } p \neq n, q = n \\ 2(\mathbf{B}_y^{(ij)}\mathbf{A})_{mn} & \text{if } p = n, q = n \end{cases} \ . \tag{12}$$

2.2.2. Remarks

Several remarks on the use of barrier methods are worth mentioning:

- As the iterative solution approaches the boundary of any constraint, the log value in Eqn (4) will tend to -ve infinity which protects the solution from violating that constraint.
- The value of $\beta(>0)$ controls the strength of the barrier. It can be shown that when $\beta \rightarrow 0$, the local minimum of F would be close to the local minimum of the original constrained problem [12].
- The initial state of A has to be in the feasible region; in our experiments, we start with the DCT matrix.
- It is possible to get into the infeasible region if we allow overshooting during the gradient descent. Hence, it is recommended to use a small learning rate and gradually decrease β until the solution is reached.

3. EXPERIMENTAL EVALUATION

The proposed guided discriminative training method was evaluated on the TIMIT database [13] to find the optimal linear feature transformation for phoneme classification. TIMIT consists of 10 utterances from each of 630 native American speakers. The training set is composed of the data from 462 speakers, while the test set is composed of the data from the remaining 168 speakers.

3.1. Feature Extraction and Acoustic Modeling

Log power spectra $\mathbf{y}$ were computed using a filterbank of $N = 24$ mel-scaled triangular filters at every 10ms over a window of 25ms of speech. They were transformed using the generalized linear transformation $\mathbf{A}$ to obtain $M = 12$ MFCCs. The final 39-dimensional acoustic vector consists of 12 MFCCs and the normalized frame energy as well as their first- and second-order time derivatives.

Each phoneme was represented by a strictly left-to-right hidden Markov model (HMM) and each HMM had 3 states and 5 Gaussian mixtures per state. A skip arc was added from the first state to the third state of each phoneme HMM to account for the fact that some phonemes occur with only two frames. In addition, there were a 1-state short-pause model and a 3-state silence model. Forty-two phoneme labels were used but they were folded into the standard 39 phoneme set before results were reported.

3.2. Experimental Procedure

The phoneme set was divided into the following five broad phoneme categories: vowels, stops, fricatives, nasals, and semi-vowels. Guided discriminative training was employed to improve the classification of each broad phoneme category separately by finding an optimal linear feature transformation using the TIMIT training set. The optimal linear transformation thus computed was used to generate MFCCs for the TIMIT test data, and phoneme classification was carried out. The amount of data of each broad phoneme category is tabulated in Table 1. The assignment of phoneme pairs to various groups in GDT were done as follows: for example, if we attempted to improve the classification performance of the vowels, all phoneme pairs consisting of at least one vowel phoneme were assigned to the discriminatory group $\mathcal{D}$; all phoneme pairs consisting of at least one short-pause or silence were assigned to the don't-care group $\mathcal{R}$; all phoneme pairs except those in $\mathcal{R}$ were assigned to the constrained group $\mathcal{C}$. Notice that $\mathcal{D} \subset \mathcal{C}$ in our case. The reason is that since the objective function only minimizes the *sum* of classification errors over *all* phoneme pairs in $\mathcal{D}$, it is possible that the optimization will sacrifice certain pairs of phonemes in $\mathcal{D}$ to get a better overall minimum. By putting all the phoneme pairs in $\mathcal{D}$ to $\mathcal{C}$, we can make sure that even if that happens, the classification degradation for those phoneme pairs cannot be worse by the user-specified degradation level δ.

Table 1. Number of phoneme tokens in the TIMT corpus.

Phoneme Category	Training Set	Test Set
vowels	59261	21552
stops	22281	7932
fricatives	23455	8355
nasals	14157	5104
semi-vowels	23664	9205
others	11656	4287
total	154474	56435

Table 2. Improvement in phoneme classification accuracy after GDT with a degradation level of 95% and 90%.

Degradation-level	Vowels	Stops	Fricatives	Nasals	Semi-vowels
(Baseline)	60.05	64.11	71.86	65.77	67.71
95%	60.54	64.93	72.38	65.62	67.96
90%	61.34	66.23	72.47	66.16	67.21

Table 3. Confusion matrix from the result of GDT with a degradation level of 90%.

Discriminatory Group	Vowels	Stops	Fricatives	Nasals	Semi-vowels	Overall
(Baseline)	60.05	64.11	71.86	65.77	67.71	66.04
Vowels	**61.34**	64.09	71.24	65.16	66.57	66.30
Stops	60.78	**66.23**	70.54	64.77	67.40	66.38
Fricatives	60.18	64.96	**72.47**	66.01	67.26	66.31
Nasals	59.97	64.13	71.98	**66.16**	66.80	65.93
Semi-vowels	60.34	64.03	72.00	65.44	**67.21**	66.14

3.3. Results and Discussions

Two degradation levels of 95% and 90% for the constrained group were investigated and the results are shown in Table 2. The confusion matrix is computed for the experiment with a degradation level of 90% and is shown in Table 3. Notice that the diagonal entries in the confusion matrix correspond to the last row of results in Table 2. The following observations are made:

- In general, when the constrained group is allowed to degrade by a greater extent, more improvement can be obtained for the discriminatory group. (See Table 2.)
- The overall classification accuracy is fairly constant over all cases; actually, there are small improvement in most cases. Thus, it seems that the performance gain by the discriminatory group usually offsets the performance degradation of the constrained group. (See Table 3.)
- Stops show the biggest gain of an absolute 2.12%, and nasals show the smallest gain of an absolute 0.39%.
- For some unknown reasons, classification performance of semi-vowels gets worse after GDT. We are looking into possible mismatch between the semi-vowels in the training data and those in the test data.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we investigate a special form of discriminative training which we call *guided discriminative training* (GDT) in the context of multi-class classification problems. We also show how GDT may be used to derive an optimal linear transformation to compute generalized MFCCs from mel-filterbank log power spectra to improve TIMIT phoneme classification. As an initial investigation of the new training method, only simple models and simple error functions were employed; modest performance gain was observed.

The GDT method is very general, and we believe that by modifying the error function, it can have many applications in speech recognition. For example, one should be able to introduce MCE type of error function to the GDT formulation so that the classification errors can be minimized directly.

5. ACKNOWLEDGEMENTS

This research is partially supported by the Research Grants Council of the Hong Kong SAR under the grant numbers CA02/03.EG04 and HKUST6201/02E.

6. REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. deSouza, and R. L. Mercer, "A New Algorithm for the Estimation of Hidden Markov Model Parameters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1988, pp. 493–496.

[2] Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell, "Distance Metric Learning with Application to Clustering with Side-information," in *Advances in Neural Information Processing Systems (NIPS 2002)*, 2002.

[3] M. J. Hunt, S. M. Richardson, D. C. Bateman, and A. Piau, "An Investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 881–884.

[4] N. Kumar and A. G. Andreou, "Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition," *Speech Communications*, vol. 26, pp. 283–297, 1998.

[5] B. H. Juang and S. Katagiri, "Discriminative Training for Minimum Error Classification," *IEEE Transaction on Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec 1992.

[6] J. S. Bridle and L. Doddi, "An Alphanet Approach to Optimising Input Transformations for Continuous Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1991, vol. 1.

[7] R. Chengalvarayan and Li Deng, "HMM-Based Speech Recognition using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 243–256, May 1997.

[8] M. Rahim and C. H. Lee, "Simultaneous ANN Feature and HMM Recognizer Design Using String-based Minimum Classification Error (MCE) Training," in *Proceedings of the International Conference on Spoken Language Processing*, 1996.

[9] A. Biem, S. Katagiri, E. McDermott, and B. H. Juang, "An Application of Discriminative Feature Extraction to Filter-Bank-Based Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 96–110, Feb 2001.

[10] B. Mak, Y. C. Tam, and Q. Li, "Discriminative Auditory Features for Robust Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, USA, 2002, vol. 1, pp. 381–384.

[11] S. M. Lee, S. H. Fang, J. W. Hung, and L. S. Lee, "Improved MFCC Feature Extraction by PCA-optimized Filter-bank for Speech Recognition," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, 2001, pp. 49–52.

[12] A. V. Fiacco and G. P. McCormick, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968.

[13] V Zue, S. Seneff, and J. Glass, "Speech Database Development at MIT: TIMIT and Beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, August 1990.