A NEW MINIMUM DIVERGENCE APPROACH TO DISCRIMINATIVE TRAINING

Jun Du[1]*, Peng Liu*[2]*, Hui Jiang*[3]*, Frank K. Soong*[2]*, Ren-Hua Wang*[1]

[1]University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]Microsoft Research Asia, Beijing, P. R. China
[3]York University, 4700 Keele Street, Toronto, Ontario M3J 1P3, Canada
unuedjwj@ustc.edu, {pengliu,frankkps}@microsoft.com, hj@cs.yorku.ca, rhw@ustc.edu.cn

ABSTRACT

We propose to use Minimum Divergence, where acoustic similarity between HMMs is characterized by Kullback-Leibler divergence, for discriminative training. The MD objective function is defined as a posterior weighted divergence measured over the whole training set. Different from our earlier work, where KLD-based acoustic similarity is pre-computed for all initial models and stays invariant in the optimization procedure, here we propose to jointly optimize the whole variable MD by adjusting HMM parameters since MD is a function of the adjusted HMM parameters. An EBW optimization method is derived to minimize the whole MD objective function. The new MD formulation is evaluated on the TIDIGITS and Switchboard databases. Experimental results show that the new MD yields relative word error rate reductions of 62.1% on TIDIGITS and 8.8% on Switchboard databases when compared with the best ML-trained systems. It is also shown the new MD consistently outperforms other discriminative training criteria, such as MPE.

Index Terms— discriminative training

1. INTRODUCTION

In the past decade, Discriminative Training (DT) has been shown to be effective in reducing word error rates of Hidden Markov Model (HMM) based Automatic Speech Recognition (ASR) systems. Some widely-used discriminative training criteria, including Maximum Mutual Information (MMI) [1, 2, 3] and Minimum Classification Error (MCE) [4], which define errors at sentence level, have been shown to be quite effective on small-vocabulary tasks [2, 4]. Recently, new criteria such as Minimum Word Error (MWE) [5] and Minimum Phone Error (MPE) [5], which focus on tuning errors at lower levels, have been proposed to improve recognition performance on Large Vocabulary Continuous Speech Recognition (LVCSR) tasks, e.g. Switchboard [5].

Because we are refining acoustic models by DT, it is reasonable to define error with high resolution, which has been proved by the success in MPE [5]. In [6], we proposed a novel approach which defines error based upon the Kullback-Leibler divergence (KLD) between the underlying HMMs [8] directly. The corresponding criterion, Minimum Divergence (MD), possesses the following advantages: 1. It is with higher resolution than any label comparison based error definition. 2. It is a general solution in dealing with any kinds of models and phone sets. As a result, MD outperforms other DT criteria on several tasks [6], and achieves better robustness in noisy conditions [7]. It is notable that in MD, the accuracy term is a function of model parameters. Hence, we can also take it into consideration in the optimization process. In this paper, we propose a integrated algorithm to update the both the posterior and the accuracy. By approximating KLD between two Gaussian mixture models (GMMs) using the KLD between the two dominating Gaussian kernels, we come up with a concrete Extended Baulm-Welch (EBW) algorithm for MD updating. The algorithm is more reasonable and efficient than that in [6], which calculates all the KLDs in advance before training. By using the algorithm, MD criterion turns out to be more concise than label comparison based criteria: 1. The accuracy, which is related with the model parameter, can also be updated; 2. No dynamic programming or other heuristic alternative is required in assessing errors. As a result, experimental results on TIDIGITS and Switchboard full-set show that MD yields further error rate reduction compared with MLE and MPE.

2. OBJECTIVE FUNCTION OF MINIMUM DIVERGENCE

From the unified viewpoint of minimum error training, the objective function can be represented as an average of the transcription accuracies of all hypotheses weighted by their posterior probabilities. For conciseness, we consider the case where there is only one training token in the formulation. In such case, criteria of minimum error training can be represented as:

$$\mathcal{F}(\lambda) = \sum_{\boldsymbol{W} \in \mathcal{M}} P_\lambda(\boldsymbol{W} \mid \boldsymbol{O}) \mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_\mathrm{r}) \qquad (1)$$

[1]The work has been done when Jun Du was visiting at Microsoft Research Asia

where λ represents the set of the model parameters; $\boldsymbol{W}_\mathrm{r}$ is the reference word sequence; $\boldsymbol{\mathcal{M}}$ is the hypotheses space or the word graph; $P_\lambda(\boldsymbol{W} \mid \boldsymbol{O})$ is the posterior probability of the hypothesis $\boldsymbol{W}$ given $\boldsymbol{O}$.

In Eq. (1), $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_\mathrm{r})$ is the accuracy measure of $\boldsymbol{W}$ given its reference $\boldsymbol{W}_\mathrm{r}$. In MWE/MPE, it is the Levinstein distance at word/phone level between hypothesis and reference. We can argue this point by advocating the final goal of discriminative training. In refining acoustic models to obtain better performance, it makes more sense to define errors in a more and more granular form of acoustic similarity. Hence, we propose to use acoustic similarity as an ultimate soft error measure. The acoustic similarity can be quantitively measured based on the KLD between underlying HMMs. Correspondingly, the MD criterion is defined as:

$$\mathcal{F}_{\mathrm{MD}}(\lambda) = -\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}} P_\lambda(\boldsymbol{W}\,|\,\boldsymbol{O}) D_\lambda(\boldsymbol{W}_\mathrm{r} \parallel \boldsymbol{W}) \qquad (2)$$

where $D_\lambda(\boldsymbol{W}_\mathrm{r} \parallel \boldsymbol{W})$ is the KLD between $\boldsymbol{W}_\mathrm{r}$ and $\boldsymbol{W}$. By adopting the MD criterion, we can refine acoustic models more directly by measuring discriminative information between a reference and other hypotheses in a more precise way. The criterion has the following advantages when compared with label comparison based error definitions: 1) It measures acoustic similarity between two underlying models, which leads to high resolution error analysis. 2) In MWE/MPE training, we need to calculate the accuracy between the reference and hypothesis sequences. The exact implementation is computationally expensive, which is avoided here. 3) Label comparison is no longer used, which alleviates the influence caused by language model and phone set. 4) $D_\lambda(\boldsymbol{W}_\mathrm{r} \parallel \boldsymbol{W})$ is directly related to model parameters λ, so it can be jointly optimized in the MD training.

3. JOINT OPTIMIZATION FOR MD

In this section, we derive an Extended Baulm-Welch (EBW) algorithm to minimize the whole MD objective function jointly based on word graphs. A major different from the original EBW algorithm for MPE is that now the accuracy term is also considered in optimization.

3.1. Weak-sense auxiliary function

To use the framework of EBW, we first introduce the following weak-sense auxiliary function:

$$\mathcal{H}(\lambda,\lambda')=\sum_{w=1}^{L}\left[g_{\lambda'}(w)\log p_\lambda(\boldsymbol{O}\mid w) - h_{\lambda'}(w)a_\lambda(w)\right] \qquad (3)$$

where w represents a word arc in a word graph with L word arcs; $p_\lambda(\boldsymbol{O} \mid w)$ is the likelihood of w; $a_\lambda(w)$ is KLD between w and the corresponding time segments in reference. $g_{\lambda'}(w)$ and $h_{\lambda'}(w)$ are defined as:

$$\begin{aligned} g_{\lambda'}(w) &= \textstyle\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}} \frac{-\partial P_\lambda(\boldsymbol{W}|\boldsymbol{O})}{\partial \log p_\lambda(\boldsymbol{O}|w)} D_\lambda(\boldsymbol{W}_\mathrm{r} \parallel \boldsymbol{W})\Big|_{\lambda=\lambda'} \\ h_{\lambda'}(w) &= \textstyle\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}} P_\lambda(\boldsymbol{W}\mid\boldsymbol{O}) \frac{\partial D_\lambda(\boldsymbol{W}_\mathrm{r}\|\boldsymbol{W})}{\partial a_\lambda(w)}\Big|_{\lambda=\lambda'} \end{aligned} \qquad (4)$$

Actually, g is the counterpart of occupation probability which is the same as that in our original formulation [6], and h is newly introduced by accuracy updating. If $D_\lambda(\boldsymbol{W}_\mathrm{r} \parallel \boldsymbol{W})$ is regarded as a constant in optimization, h does not arise. It is easy to prove that $\mathcal{H}(\lambda,\lambda')$ satisfies the conditions of the weak-sense auxiliary function in [5]:

$$\partial\mathcal{H}(\lambda,\lambda')/\partial\lambda|_{\lambda=\lambda'} = \partial\mathcal{F}_{\mathrm{MD}}(\lambda)/\partial\lambda|_{\lambda=\lambda'}$$

Note that in h, posterior is used to weight derivatives on the model KLDs, h is actually dispersing the models of all the hypotheses. That means the relative moving directions of all the models are reasonable based upon the general goal of discriminative training, However, note that $h_{\lambda'}(w) > 0$, which means the general trend is the minimizing the divergence between a hypothesis and the reference, and it is somewhat unreasonable. In this paper, we only investigate the issue of relatively dispersing the hypotheses, and leave the unreasonable part to be solved in our future research.

3.2. g and h

From the definition of g and h, we can easily obtain:

$$\begin{aligned} h_\lambda(w) &= \sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}:w\in\boldsymbol{W}} P_\lambda(\boldsymbol{W}\mid\boldsymbol{O}) \\ g_\lambda(w) &= \kappa\, h_\lambda(w)\left[d_\lambda(w) - d_\lambda^{\mathrm{avg}}\right] \end{aligned} \qquad (5)$$

where:

$$\begin{aligned} d_\lambda^{\mathrm{avg}} &= -\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}} P_\lambda(\boldsymbol{W}\mid\boldsymbol{O}) D_\lambda(\boldsymbol{W}_\mathrm{r}\parallel\boldsymbol{W}) \\ d_\lambda(w) &= -\frac{\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}:w\in\boldsymbol{W}} P_\lambda(\boldsymbol{W}\,|\,\boldsymbol{O}) D_\lambda(\boldsymbol{W}_\mathrm{r}\parallel\boldsymbol{W})}{\sum_{\boldsymbol{W}\in\boldsymbol{\mathcal{M}}:w\in\boldsymbol{W}} P_\lambda(\boldsymbol{W}\,|\,\boldsymbol{O})} \end{aligned}$$

The physical meaning of $h_\lambda(w)$ is the occupancy of the arc w, and $d_\lambda(w)$ is the average similarity of hypotheses passing through the arc w in the word graph. d_λ^{avg} is the average similarity of all hypotheses in the word graph. All these values can be calculated by Forward-Backward algorithm [6] in the word graph.

3.3. $a_\lambda(w)$

With state frame-independent assumption in HMMs, the calculation of $a_\lambda(w)$ can be decomposed down to the state level [6, 9]. Thus we obtain:

$$a_\lambda(w) = \sum_{t=b_w}^{e_w} D(s_\mathrm{r}^t \parallel s_w^t) \qquad (6)$$

where b_w and e_w represent the start and end frame of w, respectively. s_w^t and $s_{\rm r}^t$ represent a certain state at time t in arc w and the reference, respectively. Hence, $D(s_{\rm r}^t \parallel s_w^t)$ is the KLD between two output distributions of s_w^t and $s_{\rm r}^t$.

In [6], we adopt unscented transform [10] to approximate KLD of two GMMs, which has been proved quite effective. However, the approach is still too expensive to be conducted online. In this paper, to update accuracy in the training process, we adopt an alternative approach: first, given the observation, the dominant kernel with maximum posterior probability is selected for $s_{\rm r}^t$ and s_w^t. Then state-level KLD can be approximated using kernel-level KLD with the following closed-form solution:

$$D_\lambda(s_{\rm r}^t \parallel s_w^t) \approx D_\lambda(l_{\rm r}^t \parallel l_w^t) = \frac{1}{2}\Big[\log\frac{|\boldsymbol{\Sigma}_{l_w^t}|}{|\boldsymbol{\Sigma}_{l_{\rm r}^t}|} - N + \text{Trace}(\boldsymbol{\Sigma}_{l_w^t}^{-1}\boldsymbol{\Sigma}_{l_{\rm r}^t}) + (\boldsymbol{\mu}_{l_{\rm r}^t} - \boldsymbol{\mu}_{l_w^t})^\top \boldsymbol{\Sigma}_{l_w^t}^{-1}(\boldsymbol{\mu}_{l_{\rm r}^t} - \boldsymbol{\mu}_{l_w^t})\Big] \quad (7)$$

where N is the dimension of parameter vectors; $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are mean and covariance, respectively; l denotes the dominant kernel.

3.4. Statistics for EBW

Now we can come up with the statistics for EBW updating. Based upon the derivations above, we obtain:

$$\frac{\partial \mathcal{H}(\lambda,\lambda')}{\partial \lambda_{sl}} = \sum_{w=1}^{L}\sum_{t=b_w}^{e_w} \left[g_{\lambda'}(w)\frac{\partial \log p_\lambda(\boldsymbol{o}_t \,|\, s_w^t)}{\partial \lambda_{sl}} - h_{\lambda'}(w)\frac{\partial D_\lambda(s_{\rm r}^t \parallel s_w^t)}{\partial \lambda_{sl}}\right] \quad (8)$$

where λ_{sl} denotes mean or covariance for the l^{th} kernel in the s^{th} state.

The solution of the first partial derivative in Eq. (8) is the same as that of MPE [5]. Based on Eq. (7), the second partial derivative in Eq. (8) can be derived as:

$$\frac{\partial D_\lambda(s_{\rm r}^t \parallel s_w^t)}{\partial \boldsymbol{\mu}_{sl}} = \delta(s_{\rm r}^t \neq s_w^t)\delta(sl = s_w^t l_w^t)\boldsymbol{\Sigma}_{sl}^{-1}(\boldsymbol{\mu}_{sl} - \boldsymbol{\mu}_{l_{\rm r}^t})$$

$$\frac{\partial D_\lambda(s_{\rm r}^t \parallel s_w^t)}{\partial \boldsymbol{\Sigma}_{sl}} = \frac{1}{2}\delta(s_{\rm r}^t \neq s_w^t)\delta(sl = s_w^t l_w^t)\boldsymbol{\Sigma}_{sl}^{-1}\Big[\boldsymbol{I} - (\boldsymbol{\mu}_{sl} - \boldsymbol{\mu}_{l_{\rm r}^t})(\boldsymbol{\mu}_{sl} - \boldsymbol{\mu}_{l_{\rm r}^t})^\top \boldsymbol{\Sigma}_{sl}^{-1} - \boldsymbol{\Sigma}_{l_{\rm r}^t}\boldsymbol{\Sigma}_{sl}^{-1}\Big] \quad (9)$$

where $l_{\rm r}^t$ and l_w^t are the dominant kernels of $s_{\rm r}^t$ and s_w^t, respectively.

By setting $\partial\mathcal{H}(\lambda,\lambda')/\partial\lambda_{sl} = 0$ and solving it, The final statistics are gathered as follows:

$$\theta_{sl}^{\text{num}}(\boldsymbol{X}) = \sum_{w=1}^{L}\sum_{t=b_w}^{e_w}[\max(g_{\lambda'}(w),0)\gamma_{slt}(\boldsymbol{X}) + h_{\lambda'}(w)\gamma_{slt}^*(\boldsymbol{X})]$$

$$\theta_{sl}^{\text{den}}(\boldsymbol{X}) = \sum_{w=1}^{L}\sum_{t=b_w}^{e_w}\max(-g_{\lambda'}(w),0)\gamma_{slt}(\boldsymbol{X})$$

where $\boldsymbol{X} = 1, \boldsymbol{O}, \boldsymbol{O}^2$. γ_{slt} are the kernel statistics [5] related with the first partial derivative in Eq. (8). γ_{slt}^* related with the second partial derivative in Eq. (8) are defined as:

$$\begin{aligned}\gamma_{slt}^*(1) &= \delta(s_{\rm r}^t \neq s_w^t)\delta(sl = s_w^t l_w^t)\\ \gamma_{slt}^*(\boldsymbol{O}) &= \gamma_{slt}^*(1)\boldsymbol{\mu}_{l_{\rm r}^t}\\ \gamma_{slt}^*(\boldsymbol{O}^2) &= \gamma_{slt}^*(1)(\boldsymbol{\mu}_{l_{\rm r}^t}\boldsymbol{\mu}_{l_{\rm r}^t}^\top + \boldsymbol{\Sigma}_{l_{\rm r}^t})\end{aligned} \quad (10)$$

Based on the above statistics from word graphs, we adopt EBW algorithm to update HMM parameters as follows [5]:

$$\boldsymbol{\mu}_{sl} = \frac{\theta_{sl}^{\text{num}}(\boldsymbol{O}) - \theta_{sl}^{\text{den}}(\boldsymbol{O}) + D_{sl}\boldsymbol{\mu}'_{sl}}{\theta_{sl}^{\text{num}}(1) - \theta_{sl}^{\text{den}}(1) + D_{sl}}$$

$$\boldsymbol{\Sigma}_{sl} = \frac{\theta_{sl}^{\text{num}}(\boldsymbol{O}^2) - \theta_{sl}^{\text{den}}(\boldsymbol{O}^2) + D_{sl}(\boldsymbol{\Sigma}'_{sl} + \boldsymbol{\mu}'_{sl}\boldsymbol{\mu}'^\top_{sl})}{\theta_{sl}^{\text{num}}(1) - \theta_{sl}^{\text{den}}(1) + D_{sl}} - \boldsymbol{\mu}_{sl}\boldsymbol{\mu}_{sl}^\top$$

where D_{sl} [5] is the kernel-dependent smoothing factor to ensure that the objective function is concave.

4. EXPERIMENTS

In our earlier work [6], all KLD-based acoustic similarities are pre-computed in a state-level KLD matrix based on initial models and treated as constants during the MD optimization. For a typical system with thousands of tied states, it's computationally expensive. In the following experiments of this paper, we calculate the kernel-level KLD online by using Eq. (7). With kernel-level KLD, there are two advantages: 1) The KLD-based acoustic similarities actually related with HMM parameters are updated in MD optimization. 2) It almost needs no extra computation in advance.

4.1. Connected digits experiments

We first conduct connected digit string recognition experiments on the TIDIGITS database[11]. The corpus vocabulary is made of the digits 'one' to 'nine', plus 'oh' and 'zero'. All four categories of speakers, i.e., men, women, boys and girls, are used for both training and testing. The models are training using 39-dimensional MFCC features. All digits are modeled using 10-state, left-to-right whole word HMMs with 6 Gaussians per state. Because of whole word models, MPE is equivalent to MWE in this case. The acoustic scaling factor κ was set to $\frac{1}{33}$ and I-smoothing is not used on TIDIGITS. The smoothing constant E [5] was set to 2.

In figure 1, performances comparison are given on the TIDIGITS test set. 'MD1' denotes that the second partial derivative in Eq. (8) is not considered in MD formulation. 'MD2' denotes the method to optimize the whole MD objective function jointly. It is clear that MD significantly improves model accuracy compared with MPE training. The best result achieved by 'MD1' gives word error rate of 0.44%, which yields relative improvement of 62.1% and 30.2% over the ML and MPE models, respectively. Also we can observe

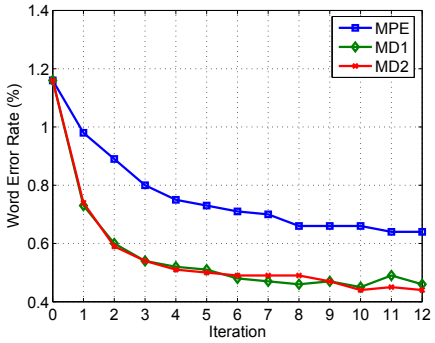

Fig. 1. *Performance comparison on TIDIGITS test set*

that 'MD2' achieves comparable performance with 'MD1'. It's reasonable because we found that the statistics defined in Eq. (10) for the second partial derivative only give a marginal benefit on the whole statistics.

4.2. LVCSR experiments

For the Switchboard task, the models are trained on the 265 hour training sets using the 39-dimensional Perceptual Linear Prediction (PLP) features with MVN (Mean and Variance Normalization) processing. Each tri-phone is modeled by a 3-state HMM. Totally, there are 6000 tied states with 16 Gaussians per state. The test set is *eval2000*. Uni-gram LM is used to generate hypotheses word graphs. Tri-gram language model is used for testing. The acoustic scaling factor κ is set to $\frac{1}{15}$. We use the NIST scoring software [12] to calculate all speech recognition results. The smoothing constant E [5] was set to 2. For MPE and MD training, I-smoothing factor τ was set to 100, and joint optimization is adopted in the latter. I-smoothing was not used for MMI because there is almost no impact to performance.

As shown in table 1, MD achieves a word error rate of 28.9%, which yields 8.8% relative improvement over the baseline. The performance is also better than MMI and MPE.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we propose to adjust HMM parameters so as to jointly optimize the whole criterion of Minimum Divergence (MD), where the accuracy term is based on model similarity and is a function of parameters now. By adopting kernel-level KLD, we obtain a closed-form EBW algorithm, which is general and more efficient than our previous solution [6]. The effectiveness of the approach is verified by experimental results, which shows that it is promising to define error and refine it based on model similarity.

As our major future work, we plan to solve the model moving direction problem mentioned in 3.1, and thus to effectively show the benefit of adjusting KLDs in optimization.

Table 1. WER (%) of different criteria on Switchboard test set

Criterion	ML	MMI	MPE	MD
WER(iter)	31.7	29.6(6)	29.1(8)	28.9(8)
Rel.Impr.	N/A	6.6%	8.2%	8.8%

6. REFERENCES

[1] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Ph.D.thesis, McGill University, 1991.

[2] R. Schluter, *Investigations on Discriminative Training Criteria*, Ph.D.thesis, Aachen University, 2000.

[3] V. Valtchev, J.J. Odell, P.C. Woodland and S.J. Young, "MMIE Training of Large Vocabulary Speech Recognition Systems", *Speech Communication*, Vol. 22, pp. 303-314.

[4] B.-H. Juang, W. Chou and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 3, pp. 257-265, May 1997.

[5] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2004.

[6] J. Du, P. Liu, F.K. Soong, J.-L. Zhou, R.-H Wang, "Minimum Divergence based Discriminative Training", *Proc. ICSLP*, pp. 2410-2413 , 2006.

[7] J. Du, P. Liu, F.K. Soong, J.-L. Zhou, R.-H Wang, "Noisy Speech Recognition Performance of Discriminative HMMs", Accepted by ISCSLP, 2006.

[8] S. Kullback and R.A. Leibler, "On Information and Sufficiency", *Ann. Math. Stat.*, Vol. 22, pp. 79-86, 1951.

[9] P. Liu, F.K. Soong, J.-L. Zhou, "Effective Estimation of Kullback-Leibler Divergence between Speech Models", *Technical Report*, Microsoft Research Asia, 2005.

[10] J. Goldberger, "An Efficient Image Similarity Measure based on Approximations of KL-Divergence between Two Gaussian Mixtures", in *Proc. International Conference on Computer Vision 2003*, pp. 370-377, Nice, France, 2003.

[11] R. G. Leonard. "A database for speaker-independent digit recognition", *Proc. ICASSP*, pp. 42.11.1-42.11.4, SanDiego, CA, March 1984.

[12] D.S. Pallett, W.M. Fisher, J.G. Fiscus, "Tools for the Analysis of Benchmark Speech Recognition Tests", *Proc. ICASSP*, pp. 97-100, 1990.