Minimum Divergence Based Discriminative Training

Jun Du[1], *Peng Liu*[2], *Frank K. Soong*[2], *Jian-Lai Zhou*[2], *Ren-Hua Wang*[1]

[1]University of Science and Technology of China, Hefei, P. R. China, 230027
[2]Microsoft Research Asia, Beijing, P. R. China, 100080

unuedjwj@ustc.edu, {pengliu,frankkps,jlzhou}@microsoft.com rhw@ustc.edu.cn

Abstract

We propose to use Minimum Divergence(MD) as a new measure of errors in discriminative training. To focus on improving discrimination between any two given acoustic models, we refine the error definition in terms of Kullback-Leibler Divergence (KLD) between them. The new measure can be regarded as a modified version of Minimum Phone Error (MPE) but with a higher resolution than just a symbol matching based criterion. Experimental recognition results show the new MD based training yields relative word error rate reductions of 57.8% and 6.1% on TIDigits and Switchboard databases, respectively, in comparing with the ML trained baseline systems. The recognition performance of MD is also shown to be consistently better than that of MPE.

Index Terms: speech recognition, discriminative training, minimum divergence, Kullback-Leibler Divergence, acoustic similarity.

1. Introduction

In the past decade, discriminative training has been shown to be effective in reducing word error rates of Hidden Markov Model (HMM) based automatic speech recognition (ASR) systems. The most widely adopted discriminative criteria, including Maximum Mutual Information (MMI) [1, 2] and Minimum Classification Error (MCE) [3], have been shown effective on small-vocabulary tasks [1, 3]. But it seems harder to obtain significant improvements on Large Vocabulary Continuous Speech Recognition (LVCSR) databases such as Switchboard task. Recently, new criteria such as Minimum Word Error (MWE) [4] and MPE [4], which are based on error measure at word or phone level, were proposed to improve recognition performance.

From a unified viewpoint of error minimization, MCE, MWE and MPE are only different in error definition. String based MCE is based upon minimizing sentence error rate and MWE on word error rate, which is more consistent with the popular metric used in evaluating ASR systems. Hence, the latter yields better word error rate, at least on the training set [4]. However, MPE performs slightly but universally better than MWE on testing set [4]. The success of MPE might be explained as follows: while we are refining acoustic models in discriminative training, it makes more sense to define errors in a more granular form of acoustic similarity. However, binary decision at phone label level is only a rough approximation of acoustic similarity. The error measure can be easily influenced by the choice of language model and phone set definition. For example, in a recognition system where whole word models are used, phone errors cannot be computed.

Therefore, we propose to use acoustic dissimilarity to define errors. Acoustic characteristics of speech units are modeled by HMMs. By measuring KLD [5] between two given HMMs, we can have a physically more meaningful assessment of their acoustic similarity. Given sufficient training data, "ideal" HMMs can be trained to represent the underlying distributions and then can be used for calculating KLDs.

Adopting KLD for defining errors, the corresponding training criterion is referred as Minimum Divergence (MD). The criterion possesses the following advantages: 1) It employs acoustic similarity for high-resolution error definition, which is directly related with acoustic model refinement; 2) Label comparison is no longer used, which alleviates the influence of chosen language model and phone set and the resultant hard binary decisions caused by label matching.

The rest of this paper is organized as follows. In section 2, we propose MD training criterion. In section 3, the algorithm for assessing KLDs on word graphs is presented. Some implementation issues are discussed in section 4. We give experimental results and our conclusions in section 5 and 6, respectively.

2. Minimum Divergence Training

In this section, in terms of minimum error training, we introduce the concept of MD criterion.

2.1. Unified view of minimum error training

In [4], various discriminative training criteria are investigated in terms of corresponding error measure, where the objective function is an average of the transcription accuracies of all hypotheses weighted by the posterior probabilities. For conciseness, we consider single utterance case:

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}}} P_\theta(\boldsymbol{W} \,|\, \boldsymbol{O}) \mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_\mathrm{r}) \tag{1}$$

where $\boldsymbol{\theta}$ represents the set of the model parameters; $\boldsymbol{O}$ is a sequence of acoustic observation vectors; $\boldsymbol{W}_\mathrm{r}$ is the reference word sequence; $\boldsymbol{\mathcal{M}}$ is the hypotheses space; $P_{\boldsymbol{\theta}}(\boldsymbol{W} \,|\, \boldsymbol{O})$ is the generalized posterior probability of the hypothesis $\boldsymbol{W}$ given $\boldsymbol{O}$, which can be formulated as:

$$P_\theta(\boldsymbol{W} \,|\, \boldsymbol{O}) = \frac{P_\theta^\kappa(\boldsymbol{O} \,|\, \boldsymbol{W}) P(\boldsymbol{W})}{\sum_{\boldsymbol{W}' \in \boldsymbol{\mathcal{M}}} P_\theta^\kappa(\boldsymbol{O} \,|\, \boldsymbol{W}') P(\boldsymbol{W}')} \tag{2}$$

where κ is the acoustic scaling factor.

In [4], the gain function $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_\mathrm{r})$ is regarded as an *accuracy* measure of $\boldsymbol{W}$ given its reference $\boldsymbol{W}_\mathrm{r}$. In MWE training, the gain function is word accuracy, which matches the commonly used evaluation metric of speech recognition perfectly. However, MPE has been shown to be more effective in testing because it provides a more precise measurement of word errors at the phone level. We

Table 1: *Comparison among minimum error criteria.* ($\boldsymbol{P_W}$ *:Phone sequence corresponding to word sequence* $\boldsymbol{W}$; LEV(,) *:Levenshtein distance between two symbol strings*; $|\cdot|$: *Number of symbols in a string.*)

Criterion	$\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$	Objective
String based MCE	$\delta(\boldsymbol{W} = \boldsymbol{W}_{\mathrm{r}})$	Sentence accuracy
MWE	$\lvert\boldsymbol{W}_{\mathrm{r}}\rvert - \mathrm{LEV}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$	Word accuracy
MPE	$\lvert\boldsymbol{P}_{\boldsymbol{W}_{\mathrm{r}}}\rvert - \mathrm{LEV}(\boldsymbol{P}_{\boldsymbol{W}}, \boldsymbol{P}_{\boldsymbol{W}_{\mathrm{r}}})$	Phone accuracy
MD	$-D(\boldsymbol{W}_{\mathrm{r}} \parallel \boldsymbol{W})$	Acoustic similarity

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 measure acoustic similarity between hypotheses instead of word accuracy. However, label matching either at word or phone level, is used in MWE or MPE training. The label matching does not relate acoustic similarity with recognition errors. The measured errors can also be strongly affected by the phone set definition and language model selection. Therefore, acoustic similarity is proposed as a finer and more direct error definition for discriminative training.

2.2. Defining error by acoustic similarity

A word sequence is acoustically characterized by a sequence of HMMs. For automatically measuring acoustic similarity between $\boldsymbol{W}$ and $\boldsymbol{W}_{\mathrm{r}}$, we adopt KLD between the corresponding HMMs:

$$\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}}) = -D(\boldsymbol{W}_{\mathrm{r}} \parallel \boldsymbol{W}) \tag{3}$$

The HMMs, when they are reasonably well trained in ML sense, can serve as succinct descriptions of data. We term this training criterion as MD. In Table 1, comparison among several minimum error criteria are tabulated.

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:

1) A strong language model can alleviate or eliminate potential problems caused by acoustically competing hypotheses in minimum error training, so usually a weaker language model is more appropriate [4]. When fine phone labels are adopted, minor acoustic difference can induce hard errors. By focusing on acoustic similarity, the above problems can be taken care of gracefully.

2) The similarity based criterion can be used in general pattern classification, where label comparison at a sub-class level may not be practical.

3. Measuring KLD in Word Graphs

In this section, we measure the KLD between a reference path and competing hypotheses in a word graph. As a word sequence is regarded as a sequence of HMMs, comparing two word sequences can be solved by measuring KLD between two sequences of HMMs.

3.1. KLD between two word sequences

Given two word sequences $\boldsymbol{W}$ and $\tilde{\boldsymbol{W}}$ without their state segmentations, we should use a state matching algorithm to measure the KLD between the corresponding HMMs [6]. With state segmentations, the calculation can be further decomposed down to the state level:

$$\begin{aligned} D(\boldsymbol{W} \parallel \tilde{\boldsymbol{W}}) &= D(\boldsymbol{s}^{1:T} \parallel \tilde{\boldsymbol{s}}^{1:T}) \\ &= \int p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}^{1:T}) \log \frac{p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}^{1:T})}{p(\boldsymbol{o}^{1:T} \mid \tilde{\boldsymbol{s}}^{1:T})} d\boldsymbol{o}^{1:T} \end{aligned} \tag{4}$$

where T is the number of frames; $\boldsymbol{o}^{1:T}$ and $\boldsymbol{s}^{1:T}$ are the observation sequence and hidden state sequence, respectively.

By assuring all observations are independent, we obtain:

$$D(\boldsymbol{s}^{1:T} \parallel \tilde{\boldsymbol{s}}^{1:T}) = \sum_{t=1}^{T} D(s^t \parallel \tilde{s}^t) \tag{5}$$

which means we can calculate KLD state by state, and sum them up.

Conventionally, each state s is characterized by a Gaussian Mixture Model (GMM): $p(\boldsymbol{o} \mid s) = \sum_{m=1}^{M_s} \omega_{sm} \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{sm}, \boldsymbol{\Sigma}_{sm})$,so the comparison is reduced to measuring KLD between two GMMs. Since there is no closed-form solution, we need to resort to the computationally intensive Monte-Carlo simulations. The unscented transform mechanism [7] has been proposed to approximate the KLD measurement of two GMMs.

Let $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a N-dimensional Gaussian distribution and h is an arbitrary $\mathbb{R}^N \rightarrow \mathbb{R}$ function, unscented transform mechanism suggests approximating the expectation of h by:

$$\int \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) h(\boldsymbol{o}) d\boldsymbol{o} \approx \frac{1}{2N} \sum_{k=1}^{2N} h(\boldsymbol{o}_k) \tag{6}$$

where $\boldsymbol{o}_k (1 \leq k \leq 2N)$ are the artificially chosen "*sigma*" points: $\boldsymbol{o}_k = \boldsymbol{\mu} + \sqrt{N\lambda_k}\boldsymbol{u}_k$, $\boldsymbol{o}_{k+N} = \boldsymbol{\mu} - \sqrt{N\lambda_k}\boldsymbol{u}_k (1 \leq k \leq N)$, where $\lambda_k, \boldsymbol{u}_k$ are the k^{th} eigenvalue and eigenvector of $\boldsymbol{\Sigma}$, respectively. Geometrically, all these "*sigma*" points are on the principal axes of $\boldsymbol{\Sigma}$. (6) is precise if h is quadratic.

Based on (6), KLD between two Gaussian mixtures is approximated by:

$$D(s \parallel \tilde{s}) \approx \frac{1}{2N} \sum_{m=1}^{M} \omega_m \sum_{k=1}^{2N} \log \frac{p(\boldsymbol{o}_{m,k} \mid s)}{p(\boldsymbol{o}_{m,k} \mid \tilde{s})} \tag{7}$$

where $\boldsymbol{o}_{m,k}$ is the k^{th} "*sigma*" point in the m^{th} Gaussian kernel of $p(\boldsymbol{o}_{m,k} \mid s)$. By plugging it into (4), we obtain the KLD between two word sequences given their state segmentations.

3.2. Word Graph based Calculation

Usually, word graph is a compact representation of large hypotheses space in speech recognition. Because the KLD between a hypothesised word sequence and the reference can be decomposed down to the frame level, we have the following word graph based representation of (1):

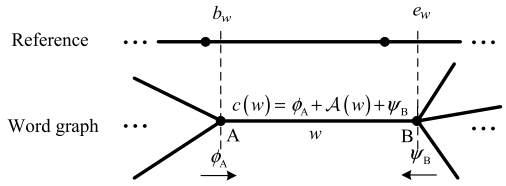

Figure 1: *Demonstration of Calculating c(w)*

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{w \in \boldsymbol{\mathcal{M}}} \sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}}: w \in \boldsymbol{W}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \,|\, \boldsymbol{O}) \mathcal{A}(w) \tag{8}$$

where $\mathcal{A}(w)$ is the gain function of word arc w. Denote b_w, e_w the start frame index and end frame index of w, we have:

$$\mathcal{A}(w) = - \sum_{t=b_w}^{e_w} D(s_w^t \,\|\, s_{\mathrm{r}}^t) \tag{9}$$

where the s_w^t and s_{r}^t represent the certain state at time t on arc w and the reference, respectively.

To update the parameters we use the Extended Baum-Welch algorithm [4] for minimum error training. The average gain of sentences passing through a word w can be represented as: [4]

$$c(w) = \frac{\sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}}, w \in \boldsymbol{W}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \,|\, \boldsymbol{O}) \mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})}{\sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}}, w \in \boldsymbol{W}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \,|\, \boldsymbol{O})} \tag{10}$$

By making use of $\mathcal{A}(w)$, we come up with an efficient forward-backward algorithm for calculating $c(w)$. The pseudo code of the algorithm is given in Table 2 for completeness, and an illustration of the measure is depicted in Figure 1. In the algorithm, given a node n, α_n and β_n denote the forward and backward likelihoods at the node; ϕ_n and ψ_n denote forward and backward average accuracy at the node; $\boldsymbol{P}(n)$ and $\boldsymbol{S}(n)$ denote the predecessor node set and successor node set of node n; $w_{m,n}$ denotes the word arc from node m to n. Given a word arc w, $B(w)$ and $E(w)$ denote the start node and end node of the arc; $P(w)$ is the likelihood of the arc. The binary operator $a \preccurlyeq b$ between two nodes a, b means that b is not an ancestor of a in the graph.

Table 2: *Forward-backward algorithm for* $c(w)$

Initialization:
- For each word arc w:
  - Calculate $\mathcal{A}(w)$
- Sort the nodes to: $n_0 \preccurlyeq n_1 \preccurlyeq \ldots \preccurlyeq n_N$

Forward:
- $\alpha_{n_0} = 1, \phi_{n_0} = 0$
- for $(i = 1; i \leq N; i++)$
  - $\alpha_{n_i} = \sum_{m \in \boldsymbol{P}(n_i)} \alpha_m \cdot P(w_{m,n_i})$
  - $\phi_{n_i} = \frac{1}{\alpha_{n_i}} \sum_{m \in \boldsymbol{P}(n_i)} [\phi_m + \mathcal{A}(w_{m,n_i})] \cdot \alpha_m \cdot P(w_{m,n_i})$

Backward:
- $\beta_{n_N} = 1, \psi_{n_N} = 0$
- for $(i = N-1; i \geq 0; i--)$
  - $\beta_{n_i} = \sum_{m \in \boldsymbol{S}(n_i)} \beta_m \cdot P(w_{n_i,m})$
  - $\psi_{n_i} = \frac{1}{\beta_{n_i}} \sum_{m \in \boldsymbol{S}(n_i)} [\psi_m + \mathcal{A}(w_{n_i,m})] \cdot \beta_m \cdot P(w_{n_i,m})$

Termination:
- For each word arc w:
  - $c(w) = \phi_{B(w)} + \mathcal{A}(w) + \psi_{E(w)}$

4. Implementation Issues

4.1. KLD precomputaion

Practically, all states in our HMM system are tied to a smaller set GMMs of several thousand states. Hence, we can precompute all KLDs between any two states to make training more efficient.

4.2. I-smoothing

I-smoothing [4] is important for minimum error training on LVCSR task. It can be regarded as using a prior of the parameters based on ML statistics. Practically, it is performed by interpolating between statistics of ML training and discriminative training:

$$\begin{aligned} \Gamma_{jm}^{'\mathrm{num}} &= \Gamma_{jm}^{\mathrm{num}} + \tau \\ \boldsymbol{\Gamma}_{jm}^{'\mathrm{num}}(\boldsymbol{o}) &= \boldsymbol{\Gamma}_{jm}^{\mathrm{num}}(\boldsymbol{o}) + \tfrac{\tau}{\Gamma_{jm}^{\mathrm{mle}}} \boldsymbol{\Gamma}_{jm}^{\mathrm{mle}}(\boldsymbol{o}) \\ \boldsymbol{\Gamma}_{jm}^{'\mathrm{num}}(\boldsymbol{o}^2) &= \boldsymbol{\Gamma}_{jm}^{\mathrm{num}}(\boldsymbol{o}^2) + \tfrac{\tau}{\Gamma_{jm}^{\mathrm{mle}}} \boldsymbol{\Gamma}_{jm}^{\mathrm{mle}}(\boldsymbol{o}^2) \end{aligned} \tag{11}$$

where Γ, $\boldsymbol{\Gamma}(\boldsymbol{o})$, and $\boldsymbol{\Gamma}(\boldsymbol{o}^2)$ denote the occupancy, first order moment and second order moment of a Gaussian kernel, respectively; the subscript jm denotes the m^{th} kernel in the j^{th} state; the superscript 'mle', 'num' indicate those in ML statistics and numerator statistics of discriminative training. I-smoothing simply means adding τ points of ML statistics to numerator statistics of discriminative training. τ is smoothing constant to control the interpolation.

5. Experiments

5.1. Connected digits experiments

we first performed experiments on TIDigits database, a connected digits recognition task [8]. 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, were used for both training and testing. The models are training using 39-dimensional MFCC features. All digits were modeled using 10-state, left-to-right whole word HMMs with 6 Gaussians per state. Because of the whole word model, MPE is equivalent to MWE now. The acoustic scaling factor κ was set to $\frac{1}{33}$ and I-smoothing was not used.

As shown in Figure 2, performance of MD achieves 57.8% relative error reduction compared with ML baseline and also outperforms MPE in all iterations.

5.2. LVCSR experiments

For the Switchboard task, the models are trained on the minitrain training sets using the 39-dimensional Perceptual Linear Prediction (PLP) features. Each tri-phone is modeled by a 3-state HMM. Totally, there are 1500 states with 12 Gaussians per state. The test set is the eval2000 set. Unigram which has been shown to be the best language model for discriminative training [1, 4] is used to

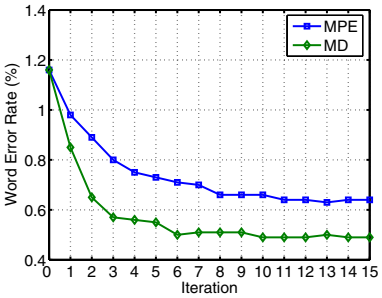

Figure 2: *Performance comparison on TIDigits*

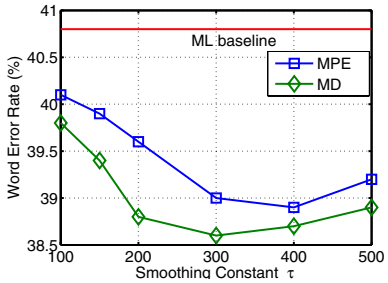

Figure 3: *Performance comparison on Switchboard with respect to smoothing constant τ after the first iteration*

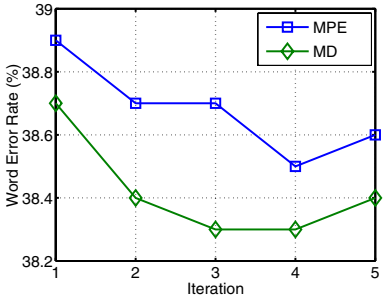

Figure 4: *Performance comparison on Switchboard with respect to iteration*

generate hypotheses word graphs in training. 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 [9] to calculate all speech recognition results. The word error rate of ML baseline is 40.8% as shown in Figure 3.

The function of τ is used to interpolate the contributions between ML and discriminative training. In minimum error training, varying τ can affect the performance significantly. We experimentally select an appropriate τ first.

With respect to τ, the recognition performance of the models after the first iteration are depicted in Figure 3. We observe that a too small or too large smoothing constant leads to lower performance as expected. Also, we found that MD outperforms MPE consistently after the first iteration, which reveals the advantage of acoustic similarity based error definition.

Note that best performances are achieved with τ around 300-400 in both MPE and MD training, here we select the value 400 for τ in the following iterations. As shown in Figure 4, performance of MD result is slightly better than MPE in all iterations. After four iterations, MD achieved 6.1% relative error reduction compared with the ML baseline, which is better than the reduction achieved by MPE. Although the improvement given by MD is not large, it is quite consistent in all experiments.

6. Conclusions and Future Work

In this paper, a new minimum divergence based discriminative training criterion, which defines errors based upon acoustic similarity, is proposed and tested. From the results we observe that by focusing on refining error measured between acoustic models, KLD based high-resolution error definition is more precise, which leads to better discriminative acoustic models and consistent recognition performance improvement.

In our future work, for more effective discriminative training on large vocabulary continuous speech recognition tasks, we will incorporate more competing hypotheses which are acoustically similar to the reference and can be obtained with a weaker language model. Symbol based error definition becomes even coarser when a weaker language model is used. By using the minimum divergence criterion, a sharpened error measure is possible and better performance is expected. Also noise robustness of MD will be investigated as parts of our future work.

7. References

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

[2] 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.

[3] 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.

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

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

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

[7] 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.

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

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