A COMPARATIVE STUDY OF DISCRIMINATIVE TRAINING USING NON-UNIFORM CRITERIA FOR CROSS-LAYER ACOUSTIC MODELING

Chao Weng, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing, Georgia Institute of Technology
75 Fifth Street NW Atlanta, GA 30308, USA
{chao.weng,juang}@ece.gatech.edu

ABSTRACT

This work focuses on a comparative study of discriminative training using non-uniform criteria for cross-layer acoustic modeling. Two kinds of discriminative training (DT) frameworks, minimum classification error like (MCE-like) and minimum phone error like (MPE-like) DT frameworks, are augmented to allow the error cost embedding at the phoneme (model) level respectively. To facilitate this comparative study, we implement both augmented DT frameworks under the same umbrella, using the error cost derived from the same cross-layer confusion matrix. Experiments on a large vocabulary task WSJ0 demonstrated the effectiveness of both DT frameworks with the formulated non-uniform error cost embedded. Several preliminary investigations on the effect of the dynamic range of error cost are also presented.

Index Terms— speech recognition, discriminative training, non-uniform error cost, cross-layer acoustic modeling

1. INTRODUCTION

Motivated by the remarkable successes of the most popular discriminative training (DT) methods, i.e., maximum mutual information (MMI)[1], minimum classification error (MCE)[2] and minimum phone/word error (MPE/MWE)[3], various contributions and several promising enhancements have been made[4]. When employing DT in many specific scenarios, however, we usually encounter a situation we call *cross-layer* acoustic modeling in that the model discrimination is often at the phoneme (model) level, while the system performance is measured at the word level (eg., WER). One issue arises from the situation is how to formulate the detriment (error cost) of the model errors to the system which is often measured at a higher level as opposed to the uniform treatment of the error cost in most current DT methods. This also gives rise to another issue, how to augment the current popular DT frameworks to be amenable for the error cost embedding. Both merit further investigation.

Since these two issues rarely invite scrutiny among the DT literature, we have explored both in our previous work. The non-uniform criteria for the DT is first initiated in [5] and then extended in [6][7][8]. As the MCE DT method aims to the direct minimization of the empirical errors with its original formulation based on the Bayes decision theory, it has been employed to demonstrate the non-uniform criteria for the *cross-layer* modeling. Meanwhile, with their approximations for incorporating the *Levenshtein distance* into the optimization, MPE-like DT methods have become popular. It would then be meaningful to compare the two: MCE with the non-uniform error cost and MPE with the non-uniform error cost.

In this work, we extend both the MPE-like and MCE-like DT frameworks to allow the error cost embedding at the model (phoneme) level, forming a comparative study of the DT using non-uniform criteria. To facilitate this comparative study, we put both under the same umbrella, using the error cost derived from the same *cross-layer* confusion matrix. Some preliminary investigations on the effect of the dynamic range of error cost are also presented. The remainder of this paper is organized as follows: The MPE-like and MCE-like DT framework are extended to allow the error cost assignment in Section 2 and Section 3 respectively. Section 4 gives an illustration of cross-layer error cost formulation. Experiments and results are reported in Section 5.

2. NON-UNIFORM ERROR MPE

2.1. MPE-like DT framework

The MPE-like DT methods, including MMI and MPE/MWE, formulate the *accuracy-based* objective functions which we want to maximize during optimization. For MPE/MWE, it is

$$F_{MPE}(\Lambda) = \sum_{r=1}^{R} \sum_{W'} \frac{P_{\Lambda}^{\alpha}(X_r|W')P^{\beta}(W')Acc(W', W_r)}{\sum_{W} P_{\Lambda}^{\alpha}(X_r|W)P^{\beta}(W)}, \quad (1)$$

where X_r and W_r are the rth training token and its label transcription, and W', W are the hypothesized transcriptions selected from the hypothesis and evidence spaces respectively. $P_{\Lambda}(X_r|W)$ and $P_{\Lambda}(W)$ denote the acoustic and language models with their scaling factors α and β respectively. $Acc(\cdot,\cdot)$ is the accuracy metric function which involves calculating the Levenshtein distance between two word sequences. The objective functions of MPE-like DT methods are optimized iteratively via the auxiliary function with the following unified form:

$$Q(\Lambda, \Lambda') = Q^{num}(\Lambda, \Lambda') - Q^{den}(\Lambda, \Lambda') + Q^{sm}(\Lambda, \Lambda'). \quad (2)$$

Here Q^{num} and Q^{den} are the auxiliary functions for the standard Baul-Welch estimation which actually are variational bounds derived from the Jensen Inequality. To compensate the negated term Q^{den} which voilates the log-concave property, the smoothing term Q^{sm} is added to guarantee the effectiveness of the whole auxiliary function in Eq. (2), for more specific forms, consult [3].

2.2. MPE extension for error cost embedding

Based on the accuracy-based form of the MPE-like objective function alone, it seems intractable to bring in the non-uniform error cost. Specifically, maximizing the auxiliary function in Eq. (2), the generic extended Baul-Welch (EBW) re-estimation formula can be

written in the following form (without I-smoothing),

$$\hat{\mu}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}(L_{jm}^{\text{num}r}(t) - L_{jm}^{\text{den}r}(t))x_t^r + D_{jm}\mu_{jm}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}(L_{jm}^{\text{num}r}(t) - L_{jm}^{\text{den}r}(t)) + D_{jm}}, \quad (3)$$

$$\hat{\Sigma}_{jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}(L_{jm}^{\text{num}r}(t) - L_{jm}^{\text{den}r}(t))x_t^r x_t^{r\text{T}} + D_{jm}G_{jm}^{\text{sm}}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}(L_{jm}^{\text{num}r}(t) - L_{jm}^{\text{den}r}(t)) + D_{jm}} - \hat{\mu}_{jm}\hat{\mu}_{jm}^{\text{T}}, \quad (4)$$

where

$$G_{jm}^{\text{sm}} = \Sigma_{jm} + \mu_{jm}\mu_{jm}^{\text{T}}, \quad (5)$$

x_t^r is the tth frame of the training token X_r. D_{jm} is the smoothing factor derived from Q^{sm} in Eq. (2). μ_{jm} and Σ_{jm} denote the Gaussian mean vector and covariance matrix for state j and mixture m of the corresponding HMM. The keys in Eq. (3) and Eq. (4) are the calculation of $L_{jm}(t)$ and the determination of the smoothing factor D_{jm}. For MMI, $L_{jm}(t)$ is the *occupancy probability* for certain state and mixture,

$$L_{jm}(t) = \gamma_{jm}(t), \quad (6)$$

which can be computed by performing the forward-backward algorithm on the decoded phoneme/word lattice and then the corresponding HMM. For MWE/MPE, $L_{jm}(t)$ has the following form:

$$L_{jm}(t) = \gamma_{jm}(t)|\overline{Acc(q)} - \overline{Acc}|. \quad (7)$$

$\overline{Acc}$ and $\overline{Acc(q)}$ are the average phoneme/word accuracy over all hypothesized transcriptions and those passing through the corresponding phoneme q respectively. They can also be approximated in a forward-backward fashion while simultaneously accumulating the corresponding phoneme q's *local accuracy*, which is defined as,

$$PhoneAcc(q_i) = \max_{q_j \in J}\left\{ \begin{array}{ll} -1 + 2e(q_i, q_j) & if\ q_i = q_j \\ -1 + e(q_i, q_j) & if\ q_i \neq q_j \end{array} \right\}, \quad (8)$$

where q_j and q_i are the reference phoneme and the hypothesis phoneme respectively. $e(q_i, q_j)$ is the relative frame overlap rate to q_j. J corresponds to the reference phoneme set at certain frame, which allows the references having boundary variations. The local accuracy defined in Eq. (8), with its value ranging from -1 to 1, utilizes the frame overlapping between the hypothesis and the reference phoneme to measure the accuracy contributions of the local phonemes. In order to take into account the non-uniform error cost of various phonemes, we modify the local phoneme accuracy in a form of the *negated error* to allow the error cost embedding, let ϵ_{ij} be the error cost of misrecognizing the phoneme q_j to q_i, as will be seen below, this modification borrows some of the essential components in the MCE formulation,

$$PhoneAcc(q_i) = \left\{ \begin{array}{ll} 0 & if\ q_i = q_j \\ -\epsilon_{ij} \cdot \ell\{d_{ij}\} & if\ q_i \neq q_j \end{array} \right\}. \quad (9)$$

Here $\ell(\cdot)$ is the sigmoid function and d_{ij} is defined as,

$$d_{ij} = -g_\Lambda(X_{t(q_i)}, q_j) + g_\Lambda(X_{t(q_i)}, q_i), \quad (10)$$

where $t(q_i)$ is the frame interval of the q_i, $g(X_{t(q_i)}, q_i)$ is the discriminant function,

$$g_\Lambda(X_{t(q_i)}, q_i) = \log P_\Lambda(X_{t(q_i)}|q_i), \quad (11)$$

3. NON-UNIFORM ERROR MCE

3.1. MCE-like DT framework

MCE-like DT methods aim at the direct minimization of the empirical error. The original MCE DT method can be summarized in the following equations,

$$g_\Lambda(X_r, W) = \log P_\Lambda^\alpha(X_r|W)P_\Lambda^\beta(W), \quad (12)$$

$$d_\Lambda(X_r) = -g_\Lambda(X_r, W_r) + \log\left[\frac{1}{|W|}\sum_{W \neq W_r}\exp[g_\Lambda(X_r, W)]\eta\right]^{\frac{1}{\eta}}. \quad (13)$$

Through this section we use the similar notations as in Eq. (1). With proper smoothing using the sigmoid function, the objective function is formulated as,

$$L_\Lambda = \sum_{r=1}^{R}\ell(d_\Lambda(X_r)). \quad (14)$$

In the original MCE methods, the model parameters are optimized using the gradient probabilistic descent (GPD)[9], in which the gradient of the objective function in Eq. (14) is approximated by a gradient at a single training sample,

$$\Lambda' = \Lambda - \mu_{GPD} \cdot \nabla\ell(d_\Lambda(X_r)) \quad r = 1, \cdots, R, \quad (15)$$

where μ is the step size. Also the GPD can be implemented in a batch mode, which is the standard gradient descent (GD) algorithm,

$$\Lambda' = \Lambda - \mu_{GD} \cdot \sum_{r=1}^{R}\nabla\ell(d_\Lambda(X_r)). \quad (16)$$

The gradient at a single training sample is given by,

$$\nabla\ell(d_\Lambda(X_r)) = \sum_{t=1}^{T_r}\gamma\ell(d_\Lambda(X_r))[1 - \ell(d_\Lambda(X_r))] (-\gamma_{jm}^{W_r}(t) + \gamma_{jm}^{W \neq W_r}(t))\frac{\partial \log \aleph_{jm}(x_t^r, \Lambda)}{\partial\Lambda}, \quad (17)$$

where $\aleph_{jm}(x_t^r, \Lambda)$ is the corresponding Gaussian, and the $\gamma_{jm}^{W_r}(t)$ and $\gamma_{jm}^{W \neq W_r}(t)$ are the *model/mixture occupancy probability* among the label and hypothesized transcriptions respectively which also can be approximated using a 0-1 indicator function determined by the Viterbi alignment.

3.2. MCE extension for error cost embedding

For the MCE-like DT methods, the error cost embedding is more intuitive. Since the original string-based MCE method manipulates the minimization of the empirical errors at the *string* level, for the error cost embedding at the phoneme (model) level, we extend the discriminant functions as follows,

$$g_\Lambda(X_r^n, q) = \log \sum_{\{W' \in W | W'(n) = q\}} P_\Lambda^\alpha(X_r|W')P_\Lambda^\beta(W'). \quad (18)$$

The summation is over those hypothesis W' with its nth phoneme identity being q. The misclassification measurement is then given by,

$$d_\Lambda(X_r^n, q_j, q_i) = -g_\Lambda(X_r, q_j) + \max_{q_i \neq q_j} g_\Lambda(X_r, q_i), \quad (19)$$

the max operation is over the hypothesis phonemes selected from the decoded lattice (word graph) at the corresponding time interval since it is prohibitive to enumerate all the other models in a large vocabulary task. Then the ultimate objective function with the error cost embedded is formulated as,

$$L_\Lambda = \sum_{r=1}^{R}\sum_{n=1}^{N_r} \ell(d_\Lambda(X_r^n, q_j, q_i)) \cdot \epsilon_{ij} \cdot 1[W_r(n) = q_j]. \quad (20)$$

With the error cost embedded, the optimization procedure will be more vulnerable without the regulation of the step size. In this work, we will use GD as the optimization method for the MCE extension, in which the step size is determined according to [10].

4. CROSS-LAYER ERROR COST FORMULATION

So far we extend both the MPE-like and MCE-like DT methods to allow the error cost embedding at the model (phoneme) level. To illustrate the cross-layer modeling, we give one possible way to formulate the non-uniform error cost ϵ_{ij}.

With the WER as the goal of minimization, to investigate how certain type of the phoneme errors would raise the word errors in a cross-layer fashion, we first define the *cross-layer confusion matrix*. Each entry C_{ij} of the *cross-layer confusion matrix* is formed in the following way: For each word in the lexicon, we pick up one arbitrary phoneme q_j from its pronunciations, then swap it with another one q_i and form a new pronunciation. The new formed pronunciation is searched over all other words among the whole lexicon, the entry C_{ij} is the number of the matched pronunciations. The rationale of deriving the error cost from the *cross-layer confusion matrix* is that the phoneme error cost with respect to word errors should be proportional to the number of words the original one may change to after its phoneme error occurs. Although a more solid formulation of the *cross-layer confusion matrix* would incorporate the uneven word prior distribution, i.e., larger mass should be put on those phonemes belonging to more frequent words accordingly, for simplicity, we will treat each word in the lexicon uniformly when forming the *confusion matrix*.

To investigate the value of C_{ij} in the real scenario, we generate a *cross-layer confusion matrix* from a lexicon, which is drawn from the SI-84 training set of WSJ database with vocabulary size of 8919. The phoneme set we use is the TIMIT 39 monophonemes set, so the size of the confusion matrix is 39×39. However we find that it would be problematic if we directly adopt the C_{ij} as the error cost due to its dynamic range in which the lowest value is 0 while some high values are beyond 200, which will be too aggressive for the parameters optimization. Meanwhile, it seems unlikely that the entry with the value of 100 truly carries 100 times the *significance* of the entry with value of 1. Obviously, we need to control the dynamic range of the ϵ_{ij}, thus the following scaling is used,

$$\epsilon_{ij} = \ln(e + \frac{C_{ij}}{\eta}), \quad (21)$$

where e guarantees the error cost is greater or equal to 1, η is to control the dynamic range of the error cost.

Here we want to emphasize the following: Deriving the error cost from the confusion matrix is just one way, but not the only way. The error cost formulation synergistically depends on the system evaluation measure, thus the cost may be introduced arbitrarily by the designer; The issue of dynamic range has a lot to do with the dispersion characteristics of the data, so it may be impossible to estimate a prior for the value of η in Eq. (21), which needs to be determined empirically.

5. EXPERIMENTS

We evaluate both kinds of the extended DT methods with the formulated cross-layer error cost embedded on the WSJ0 LVCSR database. The training corpus is the SI-84 set, which is the same as the one we generate the cross-layer confusion matrix from in Section 4, with 7133 utterances from 84 speakers and the test set is the standard Nov92 with 330 utterances from 8 speakers. The baseline system is built following the recipe (http://www.inference.phy.cam.ac.uk/kv227/htk/) for WSJ database using the Hidden Markov Model Toolkit(HTK). Cross-word tri-phone models with a total number of 2750 tied-states are trained, which are represented by 3-state strict left-to-right HMMs with each state having 8 mixture Gaussian components. The input feature is 12MFCCs + energy, and their first and second order time derivatives. The WER of the baseline system is 7.14% after 5 iterations with maximum likelihood estimation (MLE) using a standard bi-gram language model.

5.1. Non-uniform MPE-like method

For the MPE-like method with the formulated error cost embedded, we use the regular MPE (i.e., in which the local accuracy is defined in Eq. (8)) as the state-of-the-arts. While for the extended MPE-like DT methods, the η is first set to ∞ to verify the effectiveness of the extended methods, which is actually the uniform case. Then the η is set to 1, 2 and 3. For each case, we update the model parameters using EBW in 10 iterations. As mentioned in Section 2, one key parameter in EBW is the smoothing factor D_{jm}. Although D_{jm} can be theoretically determined using an upper bound derived in [11], it still can be approximated using the following heuristics as in the original MMI and MPE[12],

$$D_{jm} \approx \max\left\{E\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^{\text{denr}}(t), 2D_{jm}^{\min}\right\}, \quad (22)$$

where $D_{jm}^{\min}$ is the minimum value to guarantee the covariance matrix positive definite. I-smoothing is also employed in the experiments, in which τ is set to 200 according to [13]. The results of WER in

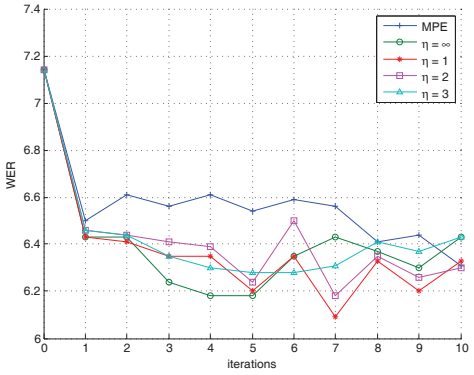

Fig. 1. WER of MPE extensions with the error cost embedded

each case during 10 iterations are shown in the Fig. 1. The extended methods almost outperform regular MPE during 10 iterations. We also list the best result of each case during 10 iterations and their relative enhancements to the baseline in table 1, which shows the non-uniform MPE with the error cost embedded in the case of $\eta = 1$ achieves the best results, about 15% relative improvement.

Table 1. Relative Improvement of Non-uniform MPE Over Baseline

Method	WER	Relative Improvement
MLE	7.14%	N/A
Regular MPE	6.31%	11.62%
Non-uniform MPE, $\eta = \infty$	6.18%	13.45%
$\eta = 1$	6.09%	14.71%
$\eta = 2$	6.18%	13.45%
$\eta = 3$	6.28%	12.04%

Table 2. Relative Improvement of Non-uniform MCE Over Baseline

Method	WER	Relative Improvement
MLE	7.14%	N/A
Non-uniform MCE, $\eta = \infty$	6.48%	9.24%
$\eta = 1$	6.39%	10.50%
$\eta = 2$	6.46%	9.52%
$\eta = 3$	6.44%	9.80%

5.2. Non-uniform MCE-like method

For the MCE-like method with the formulated error cost embedded, we implement the extended MCE methods with the values of η set to ∞,1, 2 and 3 respectively. Since the *cross-layer confusion matrix* is based on the monophonemes, the context of the label and hypothesis triphones used in both the non-uniform MPE and MCE will first be striped to obtain the value of ϵ_{ij}. For those triphones sharing the same monophone with the different context, the ϵ_{ij} is set to 1. For each case, we update the model parameters using GD in 10 iterations. According to [10], the step size μ is set utilizing a factor which can be regarded as the counterpart of D_{jm}. In this work, the factor will be set using the same heuristic in Eq. (22). As shown in Fig. 2,

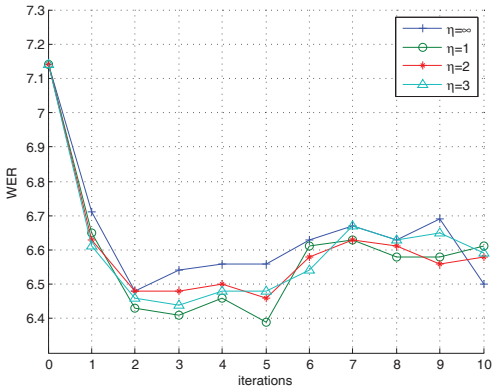

Fig. 2. WER of MCE extensions with the error cost embedded

the WER of the extended methods with the non-uniform error cost embedded is almost less than the uniform case during 10 iterations. We also list the best results of each case during 10 iterations and their relative enhancements in table 2, which shows the non-uniform MCE with the error cost embedded in the case of $\eta = 1$ achieves the best results with 10.5% relative enhancement.

5.3. Discussions

In the experiments of the MPE-like methods, although the best results of the extended methods are better than the ones of regular MPE, from the Fig. 1, there exist larger fluctuations in the extended methods when the error cost is embedded (i.e., $\eta = 1, 2, 3$). This is probably because we still use the same heuristic to set the values of D_{jm} in the EBW, which may need further modifications when the error cost is embedded. As shown in Fig. 2, the WER of the MCE-like extended methods tends to increase during the second half of the iterations. Since the η is set to a constant during the iterations, whether it can be adaptively determined in the later iterations when all the training samples are observed in the previous iteration. We leave both issues in the future work.

6. CONCLUSION

In this paper, we form a comparative study of DT using non-uniform criteria for cross-layer acoustic modeling. Two kinds of DT frameworks, MCE-like and MPE-like DT frameworks are extended to allow the cross-layer error cost embedding at the phoneme (model) level. Then the two kinds of DT frameworks are rendered under the same umbrella, with the same formulation of the non-uniform error cost derived from the cross-layer confusion matrix. Experiments are conducted to show the effectiveness of the both extended frameworks with the error cost embedded. The effects of the dynamic range of the error cost are also preliminarily investigated. We will leave more theoretical details related to the effects of the non-uniform error cost on the resultant models in the future work.

7. REFERENCES

[1] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. ICASSP1986*, 1986, pp. 49–52.

[2] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Process.*, vol. 40, pp. 3043–3054, Dec. 1992.

[3] D. Povey, *Discriminative learning for large vocabulary speech recognition*, Ph.D. thesis, Univ. of Cambridge, 2004.

[4] X. He, L. Deng, and C. Wu, "Discriminative learning in sequential pattern recognition," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 14–36, 2008.

[5] Q. Fu, D. S. Mansjur, and B.-H. Juang, "Non-uniform error criteria for automatic pattern and speech recognition," in *Proc. ICASSP2008*, 2008, pp. 1853–1856.

[6] Q. Fu, D. S. Mansjur, and B.-H. Juang, "Empirical system learning for statistical pattern recognition with non-uniform error criteria," *IEEE Trans. Signal Process.*, vol. 58, pp. 4621–4633, Oct. 2010.

[7] Q. Fu, Y. Zhao, and B.-H. Juang, "Automatic speech recognition based on non-uniform error criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, accepted for publication.

[8] C. Weng and B.-H. Juang, "Recent development of discriminative training using non-uniform criteria for cross-level acoustic modeling," in *Proc. ICASSP2011*, 2011, pp. 5332–5335.

[9] S. Katagiri, C.-H. Lee, and B.-H. Juang, "New discriminative training algorithms based on the generalized probabilistic descent method," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 1991, pp. 299–308.

[10] R. Schluter, W. Macherey, B. Muller, and H. Ney, "Comparision of discriminative training criteria and optimization methods for speech recognition," *Speech Communications*, vol. 34, pp. 287–310, 2001.

[11] T. Jebara, *Discriminative, Generative and Imitative Learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[12] M. Afify, "Extended baum-welch reestimation of gaussian mixture models based on reverse jensen inequality," in *Proc. Interspeech2005*, 2005, pp. 1113–1116.

[13] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to mpe for large scale discriminative training," in *Proc. ICASSP2007*, 2007, pp. 321–324.