A STUDY ON SOFT MARGIN ESTIMATION FOR LVCSR

Jinyu Li[1], *Zhi-Jie Yan*[2], *Chin-Hui Lee*[1] *and Ren-Hua Wang*[2]

[1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, USA, 30332
[2]iFlytek Speech Lab, University of Science and Technology of China, Hefei, P. R. China, 230027
jinyuli@ece.gatech.edu yanzhijie@ustc.edu chl@ece.gatech.edu rhw@ustc.edu.cn

ABSTRACT

We extend our previous work on soft margin estimation (SME) to large vocabulary continuous speech recognition in two aspects. The first is to use the extended Baum-Welch method to replace the conventional generalized probabilistic descent algorithm for optimization. The second is to compare SME with minimum classification error (MCE) training with the same implementation details in order to show that it is indeed the margin component in the objective function with margin-based utterance and frame selection that contributes to the success of SME. Tested on the 5k-word Wall Street Journal task, all the SME methods work better than MCE. The best SME approach achieves a relative word error rate reduction of about 19% over our best baseline performance. This enhancement can only be demonstrated because of our use of margin-based objective function and the extended Baum-Welch parameter optimization method.

Index Terms— soft margin estimation, hidden Markov model, discriminative training, extended Baum-Welch, lattice

1. INTRODUCTION

Discriminative training (DT) methods have been extensively studied to boost the automatic speech recognition (ASR) system accuracy [1-3]. The most successful methods are maximum mutual information estimation (MMIE) [1], minimum classification error (MCE) [2], and minimum word/phone error (MWE/MPE) [3]. MMIE training separates different classes by maximizing approximate posterior probabilities. On the other hand, MCE directly minimizes approximate string errors, while MWE/MPE attempts to optimize approximate word and phone error rates. If the acoustic conditions in the testing set match well with those in the training set, these DT algorithms usually achieve very good performance when tested. However, such a good match can not always be expected for most practical recognition conditions. To avoid the problem of over-fitting on the training set, regularization is achieved by using "I-smoothing" [3] in MMIE and MWE/MPE while MCE exploits a smoothing parameter in a sigmoid function for regularization [4].

Inspired by the great success of margin-based classifiers, there is a trend to incorporate the margin concept into hidden Markov model (HMM) for ASR. In contrast to the above conventional DT methods, margin-based techniques treat the generalization issue from a perspective of statistical learning theory [5]. Several attempts based on margin maximization were proposed recently and have shown some advantages over DT methods in some ASR tasks [6-10]. Among them, soft margin estimation (SME) [9] was proposed to make a direct use of the successful ideas of soft margin in support vector machines [11] to improve the generalization capability and decision feedback learning in DT to enhance model separation in the classifier design.

In [10], SME was shown to work well on the 5k-word Wall Street Journal (5k-WSJ0) task. However, two potential areas for improvement need to be addressed. The first is that the generalized probabilistic descent algorithm [12] was used for HMM parameter optimization. Although it is easy to work in a small task [9], we had a hard time getting suitable step sizes in a large vocabulary continuous speech recognition (LVCSR) task. The second is that SME improves over models initialized with maximum likelihood estimation (MLE), but fails to demonstrate its advantage over conventional DT models, such as MCE-trained models, in the same experimental configuration [10].

This study addresses the two abovementioned issues. For optimization, extended Baum-Welch (EBW) is adopted to update HMM parameters with statistics obtained from lattices. For comparison, MCE will be compared fairly with SME by sharing most of the implementation details. In addition, we build a baseline with cross-word triphone models, as opposed to the within-word models in [10], to show that SME indeed also makes significant improvements over this better baseline.

In summary, the proposed SME modification, with utterance and frame selection using EBW optimization, performs better than MCE and MLE. Above all SME with frame selection works better than SME with utterance selection because of the use of more confusion patterns. The best SME model achieves a relative word error rate (WER) reduction of 19% from our best MLE baseline.

2. SOFT MARGIN ESTIMATION

In this section, the theory of soft margin estimation is first briefly reviewed. Then we focus on how to design SME on an LVCSR task. Both utterance-based and frame-based SME methods are proposed. To make a fair comparison with MCE, these SME methods share most implementations with MCE.

2.1 Original SME Formulation

Here, we briefly introduce SME. Please refer to [9][10] for detailed discussion. According to the statistical learning theory [5], a test risk is bounded by the summation of two terms: an empirical risk (i.e., the risk on the training set) and a generalization function. The generalization function is a monotonic increasing function of Vapnik & Chervonenkis dimension, or VC dimension (VC_{dim}) [5]. Usually a classifier generalizes better with a small VC_{dim}. It can be shown that VC_{dim} is bounded by a decreasing function of the margin [5]. Hence, VC_{dim} can be reduced by increasing the margin.

This is the key idea of the margin-based method.

As analyzed, there are two targets for optimization: one is to minimize the empirical risk, and the other is to maximize the margin. These two targets are combined into a single SME objective function for minimization:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+R_{emp}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{i=1}^{N}\ell(O_i,\Lambda). \quad (1)$$

Λ denotes the set of HMM parameters, $\ell(O_i,\Lambda)$ is a loss function for utterance O_i, and N is the number of training utterances. ρ is a constant soft margin, and λ is a coefficient to balance soft margin maximization and empirical risk minimization.

The key component of SME is a proper definition of the loss function, $\ell(O_i,\Lambda)$. This loss should be related to the margin, ρ. In the original formulation of SME [9], margin is used for utterance selection with a hinge loss function. This usage will be discussed in Section 2.2. As an extension, margin-based frame selection will be discussed in Section 2.3. Margin-based utterance and frame selection allow the loss in Eq. (1) to focus on samples important to model separation instead of using all the samples.

2.2 SME with Utterance Selection

SME with utterance selection is formularized as:

$$\ell(O_i,\Lambda)=(\rho-d(O_i,\Lambda))I(O_i\in U). \quad (2)$$

I is an indicator function and selects the utterances that are in set U for the contribution of the empirical risk. $d(O_i,\Lambda)$ is the separation measure between the correct and competing candidates for O_i.

In the original formulation of SME, this utterance selection is realized via a hinge function as:

$$\ell(O_i,\Lambda)=(\rho-d(O_i,\Lambda))_+ =\begin{cases}\rho-d(O_i,\Lambda), & \text{if } \rho-d(O_i,\Lambda)>0\\ 0, & \text{otherwise}\end{cases}. \quad (3)$$

This means that if the separation measure $d(O_i,\Lambda)$ is greater than the margin, this utterance is good enough and the parameters inside it need not update. Otherwise, this utterance causes some losses and contributes to the empirical risk computation of Eq. (1).

Outlier utterances may cause trouble for model training. Hence, another loss function for utterance selection is defined as:

$$\ell(O_i,\Lambda)=\begin{cases}\rho-d(O_i,\Lambda), & \text{if } \rho>d(O_i,\Lambda)>\tau\\ 0, & \text{otherwise}\end{cases}, \quad (4)$$

where the hinge loss function in Eq. (3) is modified to have an additional threshold. The utterance that has too small value of $d(O_i,\Lambda)$ will not contribute to the loss computation because it can be an outlier.

Now, it is critical to define the separation measure $d(O_i,\Lambda)$. To make a fair comparison between SME and MCE, the following separation measure is defined:

$$d(O_i,\Lambda)=\log\frac{P_\Lambda(O_i|S_i)P(S_i)}{\sum_{\hat{S}\neq S_i\wedge\hat{S}\in G_i}P_\Lambda(O_i|\hat{S})P(\hat{S})}, \quad (5)$$

where G_i is a decoded lattice, S_i is the correct transcription for utterance O_i and $\hat{S}$ denotes the transcription of words in the decode lattice, G_i. The quantity in Eq. (5) measures the separation between the correct path and competing paths. In MCE, Eq. (5) is further embedded into a sigmoid function as in [13]. That sigmoid function can also be considered as an utterance selection function. This is because the sentences with sigmoid values close to 0 or 1 have their derivatives near 0 and will not contribute to parameter update. SME does not use a sigmoid function because there is already an utterance selection item $I(O_i\in U)$ in Eq. (2).

By plugging Eq. (5) into Eq. (3) or Eq. (4) to compute the loss in Eq. (1), SME with utterance selection is realized.

2.3 SME with Frame Selection

Since the utterances that are not selected in Eq. (3) or Eq. (4) may still have key local discriminative information from individual frames, SME with frame selection is proposed as:

$$\ell(O_i,\Lambda)=\sum_j\ell(O_{ij},\Lambda)=\sum_j\{(\rho-d(O_{ij},\Lambda))I(O_{ij}\in F_i)\}, \quad (6)$$

where O_{ij} is the jth frame for utterance O_i, and F_i is the frame set in which the frames contribute to the loss computation. SME now selects the frames that are critical to discriminative separation. We realize it with the frame posterior probability via computing the posterior probability for a word w (in the correct transcription S_i) with starting time t_{ws} and ending time t_{we}, which is got by summing the probabilities of all the lattice paths, R, in which w lies in:

$$p(w\,|\,t_{ws},t_{we},O_i)=\sum_{R\in G_i\wedge(w|t_{ws},t_{we})\in R\wedge(w|t_{ws},t_{we})\in S_i}\frac{P_\Lambda(O_i|R)P(R)}{\sum_{\hat{S}\in G_i}P_\Lambda(O_i|\hat{S})P(\hat{S})}. \quad (7)$$

The frame posterior probability is then computed by summing the posterior probabilities of all the correct words that pass time j:

$$p(S_i\,|\,O_{ij})=\sum_{w|t_{ws}<=j<=t_{we}}p(w\,|\,t_{ws},t_{we},O_i) =\sum_{\substack{w|\\ t_{ws}<=j<=t_{we}}}\sum_{R\in G_i\wedge(w|t_{ws},t_{we})\in R\wedge(w|t_{ws},t_{we})\in S_i}\frac{P_\Lambda(O_i|R)P(R)}{\sum_{\hat{S}\in G_i}P_\Lambda(O_i|\hat{S})P(\hat{S})}. \quad (8)$$

Frame selection for SME is done by comparing the frame posterior probability with the margin ρ. Using similar selection styles as in Eqs. (3) and (4), the frame loss function has the following form:

$$\ell(O_{ij},\Lambda)=\begin{cases}\rho-d(O_{ij},\Lambda), & \text{if } \rho-p(S_i\,|\,O_{ij})>0\\ 0, & \text{otherwise}\end{cases} \quad (9)$$

or

$$\ell(O_{ij},\Lambda)=\begin{cases}\rho-d(O_{ij},\Lambda), & \text{if } \rho>p(S_i\,|\,O_{ij})>\tau\\ 0, & \text{otherwise}\end{cases}. \quad (10)$$

Eqs. (9) and (10) select the frames that are critical for parameter updating. Eq. (9) focuses on confusion patterns and ignores good samples. Eq. (10) works on confusion patterns and removes the influence of noisy frames with too small posterior probabilities because they may be unreliable for parameter update due to wrong time alignment. As what will be demonstrated in the experiments section, Eq. (10) is critical for the success of frame-based SME.

The last step is to define frame level separation measure $d(O_{ij},\Lambda)$ with similar computational steps as Eq. (8):

$$d(O_{ij},\Lambda)=\log\sum_{\substack{w|\\ t_{ws}<=j<=t_{we}}}\sum_{R\in G_i\wedge(w|t_{ws},t_{we})\in R\wedge(w|t_{ws},t_{we})\in S_i}\frac{P_\Lambda(O_i|R)P(R)}{\sum_{\hat{S}\in G_i\wedge\hat{S}\neq S_i}P_\Lambda(O_i|\hat{S})P(\hat{S})}. \quad (11)$$

Similar to Eq. (5), the correct transcription is removed from the denominator in Eq. (11) because it is a measure of the correct versus the incorrect transcriptions. By plugging Eq. (11) into Eq.

(9) or (10) to compute the loss functions in Eq. (6) and Eq. (1), SME with frame selection is implemented.

2.4 Implementation with EBW

The EBW formulation with the proposed separation measure in Eqs. (5) and (11) is implemented as follows. First, an MLE model and a bigram language model (LM) were used to decode all training utterances and generate corresponding word lattices. Then a unigram was used to rescore the decoded lattices. In all the DT methods experimented in this study, a factor of 1/15 was used to scale down the acoustical model likelihood as used in the other DT studies [3][13][14]. As noted in Eqs. (5) and (11), the probabilities of the correct transcriptions are subtracted in the denominators as in [13]. Updating statistics were obtained from the lattices with a forward backward algorithm. Then, EBW was used to update the HMM parameters as in [14]. Because SME directly works on generalization, no I-smoothing was used. SME and MCE share these steps, and only differ in the definition of objective functions.

3. EXPERIMENTS

We used the 5k-WSJ0 task to evaluate the effectiveness of SME on LVCSR. The training set is the SI-84 set, with 7077 utterances from 84 speakers. All testing is conducted on the Nov92 evaluation set, with 330 utterances from 8 speakers. Baseline HMMs are trained with MLE using the HMM toolkit (HTK). The HMMs are cross-word triphone models. There were 2818 shared states obtained with a decision tree and each state observation density is modeled by an 8-mixture Gaussian mixture model. The input features were 12MFCCs + energy, and their first and second order time derivatives. A trigram LM within the 5k-WSJ0 corpus was used for decoding. The baseline WER was 5.06% for MLE models. This performance is much better than our previously baseline with within-word triphones reported in [10].

Then the MCE model was trained with the implementation in Section 2.4. The smoothing constant in the sigmoid function was set to 0.04 as in [13]. EBW was used for HMM parameters update. The WER of the MCE model was 4.60%, getting 9% relative WER reduction over the MLE baseline. This improvement percentage is similar to that reported in [13].

For the purpose of a fair comparison, all the proposed SME methods were modified on the basis of MCE implementation. This means that the implementations are similar, only the individual algorithm parts are different. SME_u and SME_uc indicate the SME models with utterance selections of Eqs. (3) and (4). SME_f and SME_fc are the SME models with the frame selections of Eqs. (9) and (10). All SME models are initiated from MLE model.

The evolutions of WERs of MCE, SME_u, and SME_fc are plotted in Figure 1. The minimum WERs of MCE, SME_u, and SME_fc were reached at iteration 12, 6, and 10, respectively. All the other methods also reached their minimum WERs within 15 iterations. The WERs of SME_fc were less than those of MCE and SME_u in every iteration.

Table 1 compares the resulting WERs and relative WER reductions of MCE and all the SME methods from MLE. All the proposed SME methods worked better than MCE, achieving about 12%-19% relative WER reduction from MLE baseline. The best was obtained by SME_fc, with the frame selection of Eq. (10).

Figure 2 compares the histograms of separation measure, *d(.)*, in Eq. (5) for MLE and SME_u models. No significant difference for the *d* values greater than 10 was observed. However, SME_u model moved the samples with *d* values less than -10 significantly to the right (resulting in bigger *d* values, which correspond to a better model separation), with 13% relative WER reduction. This demonstrates the optimization strategy of SME which focuses on confusion patterns. SME_uc achieved nearly the same result as SME_u. This shows that outlier is not a critical issue for SME with utterance selection because the utterance level information is stable. All the SME methods with utterance selection work better than MCE, showing that the utterance selection strategy in Eqs. (3) and (4) is more effective than the sigmoid function in this task.

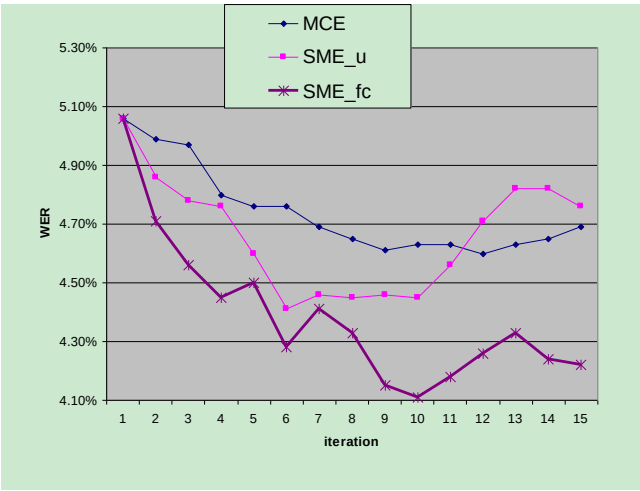

Figure 1: Evolution of testing WER for MCE, SME_u, and SME_fc models on the 5k-WSJ0 task.

Table 1: Performance comparison on the 5k-WSJ0 task

	WER	Relative Improvement
MLE	5.06%	-
MCE	4.60%	9%
SME_u	4.41%	13%
SME_uc	4.39%	13%
SME_f	4.46%	12%
SME_fc	4.11%	19%

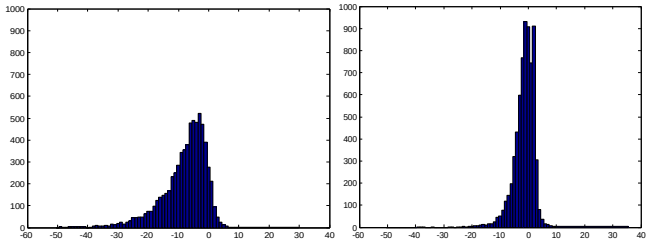

Figure 2: The histogram of the separation measure *d* in Eq. (5) of MLE (left) and SME_u (right) models on training set.

A closer look at Table 1 shows that SME_f, the SME model with hinge loss as the frame selection function, works slightly worse than SME_u. The power of using more confusion patterns did not come out clearly in this case. The reason can be well illustrated in Figure 3 by observing the histogram of the frame posterior probabilities of the MLE model (left figure). Different from the utterance separation measures in Figure 2, the distribution of posterior probabilities has two strong modes: one is around 1,

and the other is around 0. The reason that too many frames have zero posterior probability indicates that the time alignment of transcription is not precise. Therefore, given some misalignment information, the posterior probabilities are 0. These noisy frames degrade the power of using more confusion frame patterns.

As a solution, SME_fc removed the noisy frames that have too small posterior probabilities by using the loss function defined in Eq. (10). The margin and threshold of that loss function are set to be 0.8 and 0.1, respectively. Only a small amount of frames is in this range to be selected by SME_fc for parameter update. The effect is obvious: SME_fc achieved the best result, with a relative 19% WER reduction from the MLE baseline. It should be noted that the loss function in Eq. (10) is used to deal with the noisy frames for SME with frame selection because the statistics for frames is not stable. For SME with utterance selection (SME_uc), it makes little difference. As shown in the right figure of Figure 3, the SME_fc histogram of frame posterior probabilities has very few samples lying between the range of [0.1, 0.8], which is the range defined for frame selection. Most the previous samples in that range of the initial MLE model were moved to the region, [0.8, 1]. This means that SME_fc increases the model separation distances of all these confusion patterns. For the frame samples that have the posterior probability less than 0.1, SME_fc did not work on them, which is consistent with what's defined in Eq. (10).

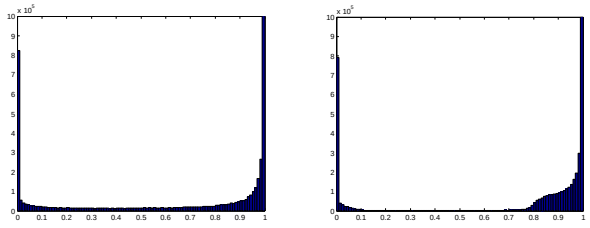

Figure 3: The histogram of the frame posterior probabilities of MLE (left) and SME_fc (right) models on training set.

4. CONCLUSION

In this study, we treated SME thoroughly on the WSJ task. As an improvement over the work reported in [10], EBW was used to update HMM parameters with the statistics collected from recognition lattices. SME realizations based on utterance selection and frame selection are realized. In contrast to SME with utterance selection, SME with frame selection uses more confusion patterns. For a fair comparison with MCE, the implementation of SME and MCE shares the same core components. Tested on the 5k-WSJ0 task, all the four proposed SME methods achieved about more than 12% relative WER reductions over MLE baseline. All SME methods also outperformed MCE. The SME model with frame selection achieved 4.11% WER, with 19% relative WER reduction from MLE, and 10% more relative WER reduction than MCE. Due to the successful frame selection strategy and powerful EBW, the result in this study is much better than the WER of 5.60% reported in our previous work [10] with only within-word triphone models. The effectiveness of SME over MCE is also well demonstrated.

Four research issues need to be addressed in the future. The first is to investigate a margin-based sample selection between the frame and utterance level. The utterance level selection discards the whole utterances with the danger of losing some helpful local discriminative information. On the other hand, the frame level selection has to deal with the noisy frames. Therefore, the unit between frame and utterance (e.g. at phone and word levels) may be a good choice because of the advantage of locality and stability. The second is to investigate whether SME can also be better than MMIE and MPE by sharing most implementations when applying the core components of SME. The third is to extend this study to feature extraction on an LVCSR task. SME already demonstrated its success in jointly optimization of features and HMM parameters on the TIDIGITS task [15]. The fourth is to apply SME to even larger LVCSR task than the 5k-WSJ0 task.

5. REFERENCES

[1] L. R. Bahl, P. F. Brown, P.V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP*, vol. 1, pp. 49-52, 1986.

[2] B. -H. Juang, W. Chou, and C. -H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 5, no. 3, pp. 257-265, 1997.

[3] D. Povey, and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," *Proc. ICASSP*, vol. 1, pp. 105-108, 2002.

[4] E. McDermott and S. Katagiri, "A derivation of minimum classification error from the theoretical classification risk using Parzen estimation," *Computer Speech and Language*, vol. 18, pp. 107-122, 2004.

[5] V. Vapnik, "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995.

[6] X. Li, H. Jiang, and C. Liu, "Large margin HMMs for speech recognition," *Proc. ICASSP*, pp. V513-V516, 2005.

[7] F. Sha and L. K. Saul, "Large margin hidden Markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems 19*, B. Sch¨olkopf, J.C. Platt, and T. Hofmann, Eds., MIT Press, 2007.

[8] D. Yu, L. Deng, X. He, and A. Acero, "Use of incrementally regulated discriminative margins in MCE training for speech recognition," *Proc. Interspeech*, pp. 2418-2421, 2006.

[9] J. Li, M. Yuan, and C. -H. Lee, "Soft margin estimation of hidden Markov model parameters," *Proc. Interspeech*, pp.2422-2425, 2006.

[10] J. Li, S. Siniscalchi, and C. -H. Lee, "Approximate test risk minimization through soft margin estimation," *Proc. ICASSP*, pp. IV 653- IV 656, 2007.

[11] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[12] S. Katagiri, B. -H. Juang and C.-H. Lee , "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proc. IEEE*, vol. 86, no. 11, pp. 2345-2373, 1998.

[13] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," *Proc. Interspeech*, pp. 2133-2136, 2005.

[14] D. Povey, "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University Engineering Dept, 2003.

[15] J. Li and C. -H. Lee, "Soft margin feature extraction for automatic speech recognition," *Proc. Interspeech*, 2007.