APPROXIMATE TEST RISK MINIMIZATION THROUGH SOFT MARGIN ESTIMATION

Jinyu Li, Sabato Marco Siniscalchi and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA. 30332 USA
{jinyuli, marco, chl}@ece.gatech.edu

ABSTRACT

In a recent study, we proposed soft margin estimation (SME) to learn parameters of continuous density hidden Markov models (HMMs). Our earlier experiments with connect digit recognition have shown that SME offers great advantages over other state-of-the-art discriminative training methods. In this paper, we illustrate SME from a perspective of statistical learning theory and show that by including a margin in formulating the SME objective function it is capable of directly minimizing the approximate test risk, while most other training methods intent to minimize only the empirical risks. We test SME on the 5k-word Wall Street Journal task, and find the proposed approach achieves a relative word error rate reduction of about 10% over our best baseline results in different experimental configurations. We believe this is the first attempt to show the effectiveness of margin-based acoustic modeling for large vocabulary continuous speech recognition. We also expect further performance improvements in the future because the approximate test risk minimization principle offers a flexible and yet rigorous framework to facilitate easy incorporation of new margin-based optimization criteria into HMM training.

Index Terms— soft margin estimation, test risk, statistical learning, lattice

1. INTRODUCTION

Recently discriminative training (DT) methods have been extensively studied to boost the automatic speech recognition (ASR) system accuracy. 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 separates different competing classes by maximizing the posterior probability. 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 in testing. However, such a good match can not always be expected for most practical recognition conditions.

From a statistical learning theory point of view [4], a test risk is bounded by the summation of two terms, an empirical risk and a generalization function. Ordinary DT methods only minimize the empirical risks, while the power to deal with possible mismatches between the training and testing conditions can often be measured by the generalization ability of the machine learning algorithms. In particular, large margin learning frameworks, such as support vector machines (SVMs) [5], have demonstrated superior generalization abilities over other conventional classifier learning algorithms. By securing a margin from decision boundaries, correct decision can still be made if the mismatched test samples fall within a tolerance region around the decision boundaries defined by the margin. Adopting the concept of enhancing margin separation, large margin estimation (LME) [6] and its variant, large relative margin estimation (LRME) [7], of HMMs have been recently proposed. In essence, LME and LRME update the models only with accurately classified samples as if the training set is indeed separable. Nevertheless, it is well known that misclassified samples are also critical for classifier learning. Recently, LRME is modified [8] to consider all the training samples, especially for the most wrongly classified sample, and move this sample in the correct decision direction. However, this modification makes the algorithm vulnerable to outliers and the idea of margin not as meaningful. In [9], a large margin algorithm for learning Gaussian mixture models (GMMs) was proposed.

In [10], we propose SME as a unified DT framework for discriminative separation, frame selection and utterance selection. Because of the incorporation of a soft margin into the optimization objective SME achieves better generalization capability and less recognition errors over LME and MCE. In this study, we illustrate the SME theory and show that the objective [10] approximates a bound of the test risk expressed as a sum of an empirical risk and a function of Vapnik & Chervonenkis dimension, or VC dimension, commonly known in statistical learning theory [4]. This is in contrast to most DT methods which attempt to minimize only the empirical risks. We also show that different choices of separation measures in loss functions lead to various approximate test risks that can be formulated as functions of string, word and phone errors and their combinations. We can therefore make use of popular losses already been used in conventional DT methods, and existing margin functions in large margin learning frameworks. This makes SME flexible and capable of incorporating new loss and margin definitions in a theoretically rigorous manner.

We evaluate SME's effectiveness on the 5k-word Wall Street Journal (5k-WSJ0) task. Two additional separation measures are proposed to take advantage of more string competition in lattices obtained in speech recognition. One is similar to the currently most successful DT algorithms by defining corresponding separation measures with statistics collected from a lattice using forward backward methods. The other is to define separation measures using word pairs appearing in a lattice. We compare the performance of the second method (SME-lattice) with those of maximum likelihood estimation (MLE) and SME with the most competing string (SME-best) proposed in [10]. Initial results on the 5k-WSJ0 task show that SME-lattice outperforms both MLE and SME-best. We expect to achieve further improvements with flexible combinations of loss and margin function definitions.

1-4244-0728-1/07/$20.00 ©2007 IEEE

Table 1: Discriminative training target function and loss function

	Optimization Objective	Loss Function l
MMIE	$\max \frac{1}{N}\sum_{i=1}^{N}\log\frac{P_\Lambda(x_i \mid S_i)P(S_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)}$	$1-\log\frac{P_\Lambda(x_i \mid S_i)P(S_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)}$
MCE	$\min \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+\exp(-\gamma f_i(X_i,\Lambda)+\theta)}$	$\frac{1}{1+\exp(-\gamma f_i(X_i,\Lambda)\ +\theta)}$,
MPE	$\max \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)RawPhoneAccuracy(\hat{S}_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)}$	$1-\frac{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)RawPhoneAccuracy(\hat{S}_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i \mid \hat{S}_i)P(\hat{S}_i)}$

2. TEST RISK BOUND

Most discriminative training methods directly minimize the risk on training set, i.e. the empirical risk, which is defined as:

$$R_{emp}(\Lambda)=\frac{1}{N}\sum_{i=1}^{N}\ell_i(X_i,\Lambda),$$

where Λ is the set of model parameters, $\ell_i(X_i,\Lambda)$ is the loss function for utterance X_i, and N is the total number of training utterances. This is shown in Table 1 listing the optimization objectives and loss functions of MMIE, MCE and MPE. S_i is the correct transcription and $\hat{S}_i$ denotes possible string sequence for utterance X_i. In MMIE and MPE, $P_\Lambda(x_i|\hat{S}_i)$ and $P(\hat{S}_i)$ are acoustic and language model scores, respectively. In MCE, f_i is a misclassification measure defined as the difference between a geometrical average of log likelihoods of competing strings and log likelihood of the correct string. γ and θ are parameters for sigmoid function. With these loss functions, these DT methods can all be considered as to minimize some empirical risks.

However, minimizing the empirical risk does not necessarily imply an optimal performance on the testing set. This can be well explained in statistical learning theory [4]. It is shown that with at least probability $1-\delta$ (δ is a small positive number) the risk on the test set, i.e. the test risk, is bounded as follows:

$$R(\Lambda)\le R_{emp}(\Lambda)+\sqrt{\frac{1}{N}\left(VC_{\dim}\left(\log(2N/VC_{\dim})+1\right)-\log\left(\frac{\delta}{4}\right)\right)}\ . \quad (1)$$

VC_{dim} is VC dimension that characterizes the classifier complexity, and can be interpreted as the maximum number of points that can be shattered by the given classification function group. This inequality shows that the test risk is bounded by the summation of two terms, the first is the empirical risk, and the second is a generalization (regularization) term which is a function of the VC dimension. Most discriminative training methods, such as MMIE, MCE, and MWE/MPE in Table 1, focus attentions on reducing the empirical risks, differing only in the choice of the loss functions, and do not consider decreasing the generalization term.

3. SOFT MARGIN ESTIMATION

In this study, we attempt to provide a theoretical perspective about SME, showing that SME directly minimizes an approximate test risk. The idea behind the choice of the loss function for SME is then illuminated. Finally we specify on the definition of separation functions. DT algorithms, such as MMIE, MCE, and MWE/MPE, can also be cast in the rigorous SME framework by defining corresponding separation functions. The solution to SME has been described in detail in [10], and will not be addressed here.

3.1 Approximate Test Risk Minimization

If we can directly minimize the right hand side of inequality (1), we can attempt to minimize the test risk. However, as a monotonic increasing function of VC_{dim}, the generalization term can not be directly minimized because it is hard to compute VC_{dim}. It can be shown that VC_{dim} is bounded by a decreasing function of a margin function [4]. Hence VC_{dim} can be reduced by increasing margin. Now, we have two targets for optimization, one is to minimize the empirical risk, and the other is to maximize margin.

We can define the SME optimization objective as follows [10]:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+R_{emp}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{i=1}^{N}\ell_i(X_i,\Lambda), \quad (2)$$

where ρ is the soft margin, λ is a coefficient to balance the soft margin maximization and the empirical risk minimization. A smaller λ corresponds to a higher penalty for the empirical risk.

Because of the relations between soft margin ρ , VC_{dim}, and the generalization term, λ/ρ has the same trend as the generalization term in Eq. (1), and can be used to approximate the generalization term. Consequently, SME directly minimizes an approximate test risk by minimizing the objective function in Eq. (2). This view distinguishes SME from both ordinary DT methods which only minimize the empirical risk $R_{emp}(\Lambda)$ and LME which only reduces the generalization term by minimizing λ/ρ in Eq. (2).

3.2 Loss Function Definition

The next issue is to define the loss function $\ell_i(X_i,\Lambda)$ for SME. As shown in Figure 1, the essence of margin-based method is to use a margin to secure some generalization in classifier learning. If the mismatch between the training and testing tests only causes a shift less than this margin in the projected space, a correct decision can still be made. So a loss happens when the separation is less than a soft margin. Therefore, the loss function can be defined as:

$$\ell_i(X_i,\Lambda)=(\rho-d_i(X_i,\Lambda))_+=\begin{cases}\rho-d_i(X_i,\Lambda), & \text{if } \rho-d_i(X_i,\Lambda)>0\\ 0, & \text{therwise}\end{cases},$$

with the SME objective function re-written as:

$$L^{SME}(\Lambda)=\frac{\lambda}{\rho}+\frac{1}{N}\sum_{i=1}^{N}(\rho-d_i(X_i,\Lambda))_+\ . \quad (3)$$

3.3 Separation Measure Definition

The third step is to define a separation (misclassification) measure, $d_i(X_i,\Lambda)$, which is a distance between correct and competing hypotheses. In [10], we defined the SME separation measure, $d_i^{SME\_utter}(X_i,\Lambda)$, which is a frame selection based log likelihood ratio of the correct and most competing string in utterance X_i. We can also define similar separation measures corresponding to MMIE, MCE, and MPE as shown in Table 2. All these measures can be put back into Eq. (3) for HMM parameter estimation.

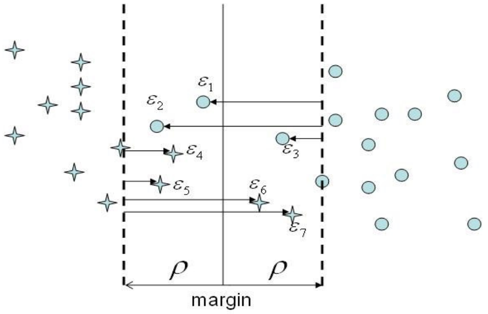

Figure 1: Soft margin estimation

Table 2: Separation measure for SME

$d_i^{SME\_utter}(X_i,\Lambda)$	$\frac{1}{n_i}\sum_j \log\left[\frac{P(X_{ij}\|S_i)}{P(X_{ij}\|\hat{S}_i)}\right] I(X_{ij}\in F_i)$
$d_i^{SME\_MMIE}(X_i,\Lambda)$	$\log\frac{P_\Lambda(x_i\|S_i)P(S_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i\|\hat{S}_i)P(\hat{S}_i)}$
$d_i^{SME\_MCE}(X_i,\Lambda)$	$1-\frac{1}{1+\exp(-\gamma f_i(X_i,\Lambda)\ +\theta)}$
$d_i^{SME\_MPE}(X_i,\Lambda)$	$\frac{\sum_{\hat{S}_i}P_\Lambda(x_i\|\hat{S}_i)P(\hat{S}_i)RawPhoneAccuracy(\hat{S}_i)}{\sum_{\hat{S}_i}P_\Lambda(x_i\|\hat{S}_i)P(\hat{S}_i)}$

4. SME ON LVCSR

The key issue for using SME in LVCSR is to define appropriate model separation measures. One method is to directly use $d_i^{SME\_utter}(X_i,\Lambda)$ in Table 2, and solve for HMM parameters by minimizing the quantity in Eq. (3). However, most successful DT methods on LVCSR use lattice to get a rich set of competing candidate information. The advantage can also be explained by the test risk bound in Eq. (1). Lattices provide more competing samples, which increase the number of training samples or a reduced generalization term, which makes a test risk bound tighter.

In the following, we provide two solutions for lattice-based separation measure definition for LVCSR.

4.1 Separation Definition in Utterance Level

The first one works in the similar way to lattice-based MMIE [11], MCE [12] and MPE [3], and define separation measures, $d_i^{SME\_MMIE}(X_i,\Lambda)$, $d_i^{SME\_MCE}(X_i,\Lambda)$, and $d_i^{SME\_MPE}(X_i,\Lambda)$, as shown in Table 2. With this kind of measures, we can easily take advantage of the optimization algorithms adopted in current lattice-based DT method, i.e. to use forward backward algorithms to get statistics from a lattice at the utterance level and then use extended Baum Welch algorithms to optimize parameters. However, because of the focus on utterance level competition, we may lose the advantage of the frame-level discrimination power in the SME separation measures as analyzed in [10].

4.2 Separation Definition in Word Level

We can also define SME separation measures at the word segment level. We first align the utterance with the correct transcription and get the timing information for every word. Next we find competing words for every word in the lattices. This is done by examining the lattice to get words falling into the time segment of the current correct transcription words. We need to set a frame overlapping threshold, so that we don't consider words with too few overlapping frames as competing words. For example, in Figure 2, we can get competing words as listed in Table 3. Finally we figure out the frames that are overlapped between the correct word and competing word. For each overlapping word pair, we denote the number of overlapped frames as n_o, the jth overlapping frame as X_{oj}, the overlapped frame set as F_o, and the target and competing words *as* W_{target} and W_{comp}. A word level separation can be defined as:

$$d_o^{SME\_word}(X_i,\Lambda)=\frac{1}{n_o}\sum_j \log\left[\frac{P(X_{oj}|W_{\text{target}})}{P(X_{oj}|W_{\text{comp}})}\right] I(X_{oj}\in F_o).\ (4)$$

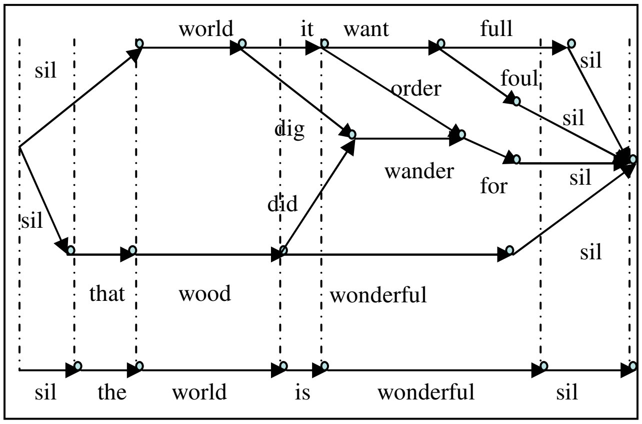

Figure 2: Lattice example. The top lattice is obtained in decoding, and the bottom is the corresponding utterance transcription.

Table 3: Correct and competing words for lattice example

Correct Word	Competing Words
the	That
world	wood, it, dig
is	it, dig, did, wonderful
wonderful	want, full, foul, order, dig, did, wander, for

The word level separation measure, $d_o^{SME-word}(X_i,\Lambda)$, is better than utterance level measure, $d_i^{SME\_utter}(X_i,\Lambda)$, because with the usage of word pairs in lattices, it can employ much more confusion patterns than $d_i^{SME\_utter}(X_i,\Lambda)$, which only use the correct and the most competitive strings. For usage in SME, $d_o^{SME-word}(X_i,\Lambda)$ may also have an advantage over separation measures defined in Section 4.1, which have only one value for each utterance. In SME we will plug this separation value into Eq. (3), and the utterances with values greater than the value of the margin will not contribute to parameter optimization. However in some cases, there may be some word pairs in lattices that still have distances less than the value of the margin. The word level $d_o^{SME\_word}(X_i,\Lambda)$ makes use of those word pairs to get more confusion patterns.

5. EXPERIMENT

We used the 5k-WSJ0 task to evaluate the effectiveness of SME on LVCSR. The training material is SI-84 set, with 7077 utterances from 84 speakers. The testing material is Nov92 evaluation set, with 330 utterances from 8 speakers. Baseline HMMs are trained with MLE using HTK. The HMMs are within-word triphone models. There are totally 2329 shared states obtained with a decision tree and each state observation density is modeled by an 8-mixture GMM. The input features are 12MFCCs + energy, and their first and second order time derivatives. The bigram and trigram language models (LMs) within the 5k-WSJ0 vocabulary were used for decoding. The baseline WERs are 8.41% with bigram LM and 6.13% with trigram LM, respectively. Other research site baselines may be better than our baseline by using different configurations. In this study, because we don't have access to those baseline configurations, we only improve over our best available setup. Our HTK-trained baselines are comparable with the HTK-trained results reported in [13], and recent results in [14]. We expect to improve over higher baseline results as well.

We used the bigram LM to get seed lattices for the training utterances. Seed lattices were generated only once. At every iteration, the recently updated HMMs were incorporated into generating new lattices by using seed lattices as decoding word graphs. After that, SME was used to update HMM parameters. This greatly improves the lattice generation speed. Two SME methods are used here. One, denoted by SME_utter, is based on separation measure, $d_i^{SME\_utter}(X_i,\Lambda)$, in Table 2. The other, denoted by SME_word, is based on the word level separation measure, $d_o^{SME\_word}(X_i,\Lambda)$, defined in Eq. (4).

In Table 4, the WERs obtained with these two SME methods and MLE are compared with. Both SME methods achieved better WERs than MLE. By taking advantage of lattices, SME_word is better than SME_utter because SME_word offers much more confusion patterns than SME_utter, which only uses most competitive string in an utterance. SME_word decreased WERs significantly from MLE, with the relative WER reductions of 12% for bigram LM and 9% for trigram LM, respectively.

Table 4: Performance comparison on the 5k-WSJ0 task

WER	Bigram	Trigram
MLE	8.41%	6.13%
SME_utter	8.14%	5.94%
SME_word	7.38%	5.60%

6. CONCLUSION

From the view of statistical learning theory, we show that SME can minimize the approximate risk on the test set. This is in contrast with most discriminative training methods, which only minimize the risk on the training test. The choice of various loss functions is illuminated and different kinds of separation measures are defined under a unified SME framework. We apply SME to LVCSR by defining separation measures at both the utterance and word levels. Because of the usage of much more confusion patterns in lattices, SME with word level separation measures performs better than SME with utterance level separation measures. Tested on the 5k-WSJ0 task, SME with word level separation measures achieves about 10% relative WER reduction over our best MLE baselines.

This paper is our first study to apply SME to LVCSR. We are now working on many related research issues. The first is to design a good optimization method than the generalized probabilistic descent algorithm which is slow in convergence. More efficient parameter update methods will be explored later. Second, we will apply separation measures defined in Section 4.1 on SME. Third, we will explore more elaborated definitions of margin functions to tightly couple it with the definition of the empirical risks.

7. ACKNOWLEDGEMENT

This work was partially supported by the NSF grant, IIS-04-27113, and the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022.

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

[5] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, 1998.

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

[7] C. Liu, H. Jiang, and X. Li, "Discriminative Training of CDHMMS for Maximum Relative Separation Margin," *Proc. ICASSP*, pp. I101-I104, 2005.

[8] C. Liu, H. Jiang, and L. Rigazio, "Recent Improvement on Maximum Relative Margin Estimation of HMMs for Speech Recognition," *Proc. ICASSP*, pp. I269-I272, 2006.

[9] F. Sha, and L. K. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," *Proc. ICASSP,* pp. I265-I268, 2006.

[10] J. Li, M. Yuan, and C. -H. Lee, "Soft Margin Estimation of Hidden Markov Model Parameters," *Proc. Interspeech*, pp. 2422-2425, 2006.

[11] V. Valtchev, J. Odell, P. Woodland, and S. Young, "MMIE Training of Large Vocabulary Recognition Systems," *Speech Communication*, vol. 22, no. 4, pp. 303-314, 1997.

[12] 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.

[13] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large Vocabulary Continuous Speech Recognition Using HTK," *Proc. ICASSP*, pp. II125-II128, 1994.

[14] Q. Fu, A. M. Daniel, B. -H. Juang, J. L. Zhou, and F. K. Soong, "Generalization of the Minimum Classification Error (MCE) Training Based on Maximizing Generalized Posterior Probability (GPP)," *Proc. Interspeech*, pp. 681-684, 2006.