On a Generalization of Margin-Based Discriminative Training to Robust Speech Recognition

Jinyu Li and Chin-Hui Lee

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

Abstract

Recently, there have been intensive studies of margin-based learning for automatic speech recognition (ASR). It is our believe that by securing a margin from the decision boundaries to the training samples, a correct decision can still be made if the mismatches between testing and training samples are well within the tolerance region specified by the margin. This nice property should be effective for robust ASR, where the testing condition is different from those in training. In this paper, we report on experiment results with soft margin estimation (SME) on the Aurora2 task and show that SME is very effective under clean training with more than 50% relative word error reductions in the clean, 20db, and 15db testing conditions, and still gives a slight improvement over conventional multi-condition training approaches. This demonstrates that the margin in SME can equip recognizers with a nice generalization property under adverse conditions.

Index Terms: soft margin estimation, hidden Markov model, robust speech recognition

1. Introduction

Despite many years of research effort robust automatic speech recognition (ASR) remains a challenging problem. The main difficulty arises from many possible types of signal distortions, such as additive and convolutive noises, and they are often mixed and typically not easy to characterize analytically. As a result, the speech recognizer trained using clean speech often degrades its performance significantly when used under noisy situations if no distortion compensation is applied [1].

Different techniques have been proposed for environment robustness over the past three decades. There are three main areas of focus. In the signal domain, the testing speech signal can be enhanced before processing (e.g., spectral subtraction [2]). Moreover in the feature domain, the distorted acoustic features can be normalized or compensated to match training features (e.g., cepstral mean normalization [3], and stereo-based piecewise linear compensation for environments [4]). Furthermore in the model domain, the original trained model can be adjusted to a model set that matches the testing environment (e.g., maximum likelihood linear regression [5], and maximum likelihood stochastic matching [6]).

In contrast to the above methods, margin-based learning may provide a set of models with generalization capabilities to deal with noise robustness without actual compensation at operating time. Inspired by the success of margin-based classifiers, there is a new trend to apply the margin concept to training hidden Markov models (HMMs). Several attempts based on margin maximization were proposed recently for discriminative training of acoustic models for ASR. They are: large margin estimation (LME) [7], large margin hidden Markov models (LM-HMMs) [8], and soft margin estimation (SME) [9]. The formulation of margin-based methods allows some mismatch between the training and testing conditions. By securing a margin from the decision boundaries to the training samples, a correct decision can still be made if the mismatches between the testing and training samples are smaller than the value of the margin. Although this nice property of margin-based methods is quite desirable, we are not aware of any previously reported work on robust ASR with margin-trained HMMs. We study discriminative training (DT) methods, such as minimum classification error (MCE) [10] and SME training, and investigate if they generalize well to adverse conditions without applying any special compensation techniques.

The rest of the paper is organized as follows. In Section 2, we review theory of SME and MCE. Then both methods are evaluated for generalization on the Aurora2 task [11] in Section 3. It is concluded that SME is more effective than MCE to handle the mismatch between the training and testing conditions under clean training in all signal-to-noise-ratio (SNR) cases. As for multi-condition training SME slightly improved over MCE.

2. Discriminative Training Methods

Theory of SME is first reviewed in this following. MCE will then be briefly described. Both methods will be evaluated on the Aurora2 connected-digit recognition task.

2.1 Soft Margin Estimation (SME)

SME [9] originates from statistical learning [12]. It is shown that the test risk is bounded by the summation of two terms, the first is an empirical risk (i.e., the risk on the training set), and the second is a generalization term which is bounded by a decreasing function of the margin [12]. Hence, SME has two targets for optimization. The first is to minimize the empirical risk. The other is to maximize the margin, which is related to classifier generalization. These two objectives are combined into a single function for minimization as follows [9]:

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

where Λ denotes the set of HMM parameters, $\ell(O_i, \Lambda)$ is a loss function for utterance O_i, N is the number of training utterances. 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.

The loss function is defined with the help of a hinge loss function ($(x)_+ = max(x,0)$) as

$$\ell(O_i, \Lambda) = [\rho - d(O_i, \Lambda)]_+ = \begin{cases} \rho - d(O_i, \Lambda), & \text{if } \rho - d(O_i, \Lambda) > 0 \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

with the separation measure d defined as

$$d(O_i, \Lambda) = \frac{1}{n_i} \sum_j \log \left[\frac{P_\Lambda(O_{ij} | S_i)}{P_\Lambda(O_{ij} | \hat{S}_i)} \right] I(O_{ij} \in F_i). \quad (3)$$

F_i is the frame set in which the frames have different labels in the competing strings. n_i is the number of frames in F_i. $I(.)$ is an indicator function, and O_{ij} is the j-th frame of utterance O_i. $P_\Lambda(O_{ij} | S_i)$ and $P_\Lambda(O_{ij} | \hat{S}_i)$ are the likelihood scores for the target string S_i and the most competing string $\hat{S}_i$.

Plugging Eq. (3) into Eq. (2), the final objective function to minimize for SME is:

$$\begin{aligned} L^{SME}(\rho, \Lambda) &= \frac{\lambda}{\rho} + \frac{1}{N} \sum_{i=1}^{N} [\rho - d(O_i, \Lambda)]_+ \\ &= \frac{\lambda}{\rho} + \frac{1}{N} \sum_{i=1}^{N} [\rho - d(O_i, \Lambda)] I[\rho - d(O_i, \Lambda) > 0] \end{aligned} . (4)$$

To solve Eq. (4), the indicator function I is approximated with a sigmoid function. Then Eq. (4) becomes

$$L^{SME}(\rho, \Lambda) = \frac{\lambda}{\rho} + \frac{1}{N} \sum_{i=1}^{N} (\rho - d(O_i, \Lambda)) \frac{1}{1 + \exp(-\gamma(\rho - d(O_i, \Lambda)))}, (5)$$

where γ is a smoothing parameter for the sigmoid function. The quantity in Eq. (5) is a smooth function of the soft margin ρ and the HMM parameters Λ. Therefore, they can be solved iteratively using the generalized probabilistic descent (GPD) algorithm [13], with η_t and κ_t as step sizes:

$$\begin{cases} \Lambda_{t+1} = \Lambda_t - \eta_t \nabla L^{SME}(\rho, \Lambda) |_{\Lambda = \Lambda_t} \\ \rho_{t+1} = \rho_t - \kappa_t \nabla L^{SME}(\rho, \Lambda) |_{\rho = \rho_t} \end{cases} .$$

2.2 Minimum Classification Error (MCE)

MCE minimizes a smoothed 0-1 loss function [10]:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{1}{1 + \exp(-\gamma h(O_i, \Lambda) + \theta)},$$

where h is a misclassification measure for utterance O_i defined as the difference between the geometrical average of the log likelihoods of competing strings and the log likelihood of the correct string. γ and θ are parameters for a sigmoid function.

For relative small tasks, GPD was often used for HMM parameter optimization. For large vocabulary continuous speech recognition tasks, extended Baum-Welch (EBW) [14] is adopted to update HMM parameters.

3. Experiment

The generalization issue for the above DT methods was evaluated on the standard Aurora 2 task of recognizing digit strings in noise and channel-distorted environments. The clean training set and multi-condition training set, which consist of 8440 clean utterances and multi-condition utterances, individually, were used to train the baseline maximum likelihood estimation (MLE) HMMs. The test material consists of three sets of distorted utterances. The data in set-A and set-B consist of eight different types of additive noise, while set-C contains two different types of noise plus additional channel distortion. Each type of noise is added into a subset of clean speech utterances, with seven different levels of SNRs. This generates seven SNR-specific subgroups, namely clean, 20db, 15db, 10db, 5db, 0db, and -5db SNRs, of testing sets for each specified noise type. The baseline experiment configuration follows the standard script provided by ETSI [11], including the simple "backend" of HMMs trained using HTK. The acoustic features are 13-dimension MFCCs, appended by their first- and second-order time derivatives. The baseline clean-trained and multi-condition-trained HMMs got 60.06% and 86.39% word accuracy (Acc), separately.

For a fair comparison, both SME and MCE were trained with similar implementations and differed only in the individual algorithm parts. Because the Aurora2 task is a connected-digit task, we need not use lattice for competing strings and EBW for parameter optimization, GPD was used to train HMM parameters and N-Best lists were used to construct competing strings. SME used only one competing string while MCE used 5 competing strings. If only one competing string was used for MCE, a worse performance was obtained. Therefore, we just reported the best performance for MCE with 5 competing strings.

3.1 Clean Training Condition

Usually, DT methods are only reported to work on the multi-style training set or on the de-noised testing condition combined with other noise robustness techniques (e.g., [15]). In this section, we investigate whether DT methods can work well in a mismatched condition (i.e., clean training case) without using other noise robust technologies.

Table 1 (all the tables are in the last page of this paper) lists the detailed test accuracies for MLE and DT methods (MCE and SME with different values of the balance coefficient λ) trained with clean data. The average digit accuracy was evaluated by averaging the accuracies on the subgroups with the SNRs from 0db to 20db, as described by the ETSI standard [11]. Table 2 shows the relative WER reductions for MCE and SME from the baseline. In average, SME with different balance coefficients achieved about 17%-29% relative WER reductions from MLE while MCE obtained only 0.2% relative WER reduction due to poor performance in low SNR conditions. Examining the results in detail, we see that the characteristics of individual recognition accuracies for different SNR subgroups are very different for these two DT methods.

For each SNR subgroup (column) in Tables 1-3, the best performance is shown in bold font. In all cases in Tables 1-2, SME outperformed MCE. It is interesting to note that for the clean testing subgroup, both SME and MCE got comparable accuracies, with the relative WER reductions ranging from 43% to 57%. This implies that under matched conditions, both DT methods performed similarly. The major difference occurred in the mismatched testing subgroups. In 20db and 15db SNR conditions that are not severely distorted from the clean training conditions, MCE got 23% and 12% relative WER reductions. In contrast, all the SME methods achieved at least 42% relative WER reductions in the 20db and 15db SNR cases, with the best performance around 60% WER reduction. In the 10db case, MCE got less than 5% relative WER reductions while most SME can still get more than 40%. In the 5db, 0db, and -5db SNR scenarios, which are severely distorted conditions, MCE can get even worse performance than the MLE baseline. In contrast, SME still obtained satisfactory relative WER reductions in all cases.

The above observations show that all these discriminative training methods have no big difference in matched testing conditions on the Aurora2 task. Big difference exists in the mismatched testing conditions. Because of the margin, SME greatly improved the generalization ability, allowing the classifier to make a correct decision as long as the testing samples deviate within the margin from the training samples.

Examining the results of Table 2 in detail, we see that the best relative WER reductions for clean, 20db, and 15db SNR testing cases are 57%, 60%, and 59%. This demonstrates the effectiveness of the margin in the cases of 20db and 15db SNR testing conditions which are not as severely distorted from clean training condition, the margin can easily cover the distortion. However, for the 10db, 5db, 0db, and -5db cases, the relative WER reductions keep decreasing since these conditions are increasingly pulled away from clean training and the distortions cannot be easily covered by a margin.

SME, with different balance coefficients λ, affects the performance in different testing conditions. In the cleaning testing case, SME with the smallest λ (λ=10) gives the best accuracy, since it is a matched testing and the classifier with a focus on empirical risk minimization works the best. With the testing SNR decreasing, larger λ is required to give more weights to margin maximization which in turn gives better generalization. As a result, SME with λ=50 works best in the 20db SNR case, while SME with λ=100 gets best accuracies in the 15db and 10db SNR cases. SME with λ=200, 300, and 400 obtain the most word error reduction in the 5db, 0db, and -5db SNR testing conditions. We can easily see the trend of best SME with respect to the coefficient.

The original generalization property of margin-based classifiers in statistical learning theory [12] requires the training and testing samples to be from the identically independent distributions (i.i.d.). The results here show that SME performs very well even if the training and testing distributions are very different, which means that SME may have even better generalization property. We believe this is because the formulation of SME can push training samples away from the decision boundary with a distance of the margin. That margin in turn allows correct decision be made as long as the testing samples are distorted from training samples within a distance less than the value of the margin.

3.2 Multi-condition Training Condition

Table 3 lists test results for MLE, MCE, and SME using multi-condition training data. Both DT methods obtained similar performance, with around 10% relative WER reduction, which is similar to the relative WER reduction reported in [15] when MCE was only applied to multi-condition trained HMMs without combining with other methods. Here, only SME results with λ=10 is given for comparison. SME with other balance coefficient gives similar results and are omitted here.

The improvement of SME in multi-condition training is not as impressive as that in the clean training case. The possible reason is given in the following. For model trained from the clean data, the accuracy is as high as 99% on the clean testing data. From Eq. (1), we can see that classifier learning balances empirical risk minimization and margin maximization. Because the empirical risk (i.e., the risk on the training set) is already very small, the focus of classifier learning is to maximize the margin. The resulted large margin in turn gives better generalization for the classifier, making it performs very well in mismatched testing conditions although the classifier is trained only with clean data. In contrast for models trained from multi-condition data, the accuracy is only around 86% on the multi-condition test data. The classifier training has to care about both the empirical minimization and the margin maximization. As a result, the margin cannot play a significant role to contribute to significantly improving the generalization capabilities of SME-trained models.

4. Conclusion

We have evaluated the generalization issues of SME and MCE in this study. Multi-condition testing with both clean and multi-condition training is investigated on the Aurora2 task. In the clean training case, SME achieves an overall average of 29% relative WER reductions while MCE gets less than 1% relative WER reductions. Although both methods perform similarly when testing with clean utterances, SME outperforms MCE significantly in the testing utterances with SNRs ranging from -5db to 20db. In those mismatched conditions, the margin in SME contributes to classifier generalization and results in great performance improvements for SME. In multi-condition training, SME is slightly better than MCE since in this case the focus of classifier learning is more on minimizing the empirical risk instead of maximizing the margin for generalization. We hope the observations in this study can further deepen the research of generalization property of margin-based classification methods.

This paper only presents our initial study, we are now working on a number of related research issues. First, current evaluation of these DT methods is on Aurora2, which is a connected-digit task. We may extend the evaluation to a larger task, such as Aurora4. Second, SME may be combined with other robust ASR methods as in [15] to further improve ASR performance.

5. References

[1] Gong, Y., “Speech recognition in noisy environments: a survey,” *Speech Communication*, vol. 16, no. 3, pp. 261-291, 1995.

[2] Boll, S. F., "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113-120, 1979.

[3] Atal, B., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identifition and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304-1312, 1974.

[4] Deng, L., Acero, A., Plumpe, M., and Huang, X., "Large vocabulary speech recognition under adverse acoustic environments," *Proc. Interspeech*, 2000.

[5] Leggetter, C. J. and Woodland, P. C., "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, 1995.

[6] Sankar, A. and Lee, C.-H., "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 4, no. 3, pp. 190-202, 1996.

[7] Jiang, H., Li, X., and Liu, C., "Large margin hidden Markov models for speech recognition," *IEEE Trans. On Audio, Speech, and Language Proc.*, vol. 14, pp. 1584-1595, 2006.

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

[9] Li, J., Yuan, M., and Lee, C. -H., "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 8, pp. 2393- 2404, 2007.

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

[11] Hirsch, H. G. and Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. ISCA ITRW ASR*, 2000.

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

[13] Katagiri, S., Juang, B. -H., and Lee, C.-H., "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.

[14] Normandin, Y., *Hidden Markov models, maximum mutual information estimation, and the speech recognition problem*. Ph.D. thesis, McGill University, 1991.

[15] Wu, J. and Huo, Q., "An environment-compensated minimum classification error training approach based on stochastic vector mapping," *IEEE Trans. On Audio, Speech, and Language Proc.*, vol. 14, no. 6, pp. 2147 – 2155, 2006.

Table 1: Detailed test accuracies for MLE, MCE, and SME with different balance coefficient λ using clean training data.

Word Acc	clean	20db	15db	10db	5db	0db	-5db	Avg.
MLE	99.03	94.07	85.04	65.52	38.61	17.09	8.53	60.06
MCE	99.50	95.43	86.88	66.95	37.29	14.27	6.58	60.16
SME ($\lambda = 10$)	**99.58**	97.00	91.39	75.73	48.61	21.59	8.53	66.86
SME ($\lambda = 50$)	99.56	**97.61**	93.42	80.74	55.32	27.29	11.01	70.88
SME ($\lambda = 100$)	99.54	97.60	**93.85**	**81.98**	56.59	28.19	12.28	71.64
SME ($\lambda = 200$)	99.49	97.55	93.76	81.94	**56.96**	28.67	13.14	**71.78**
SME ($\lambda = 300$)	99.46	97.41	93.55	81.43	56.54	**28.86**	13.33	71.56
SME ($\lambda = 400$)	99.45	97.34	93.49	81.57	56.32	28.68	**13.44**	71.48

Table 2: Relative WER reductions for MCE, and SME from MLE baseline using clean training data.

Rel. WER red.	clean	20db	15db	10db	5db	0db	-5db	Avg.
MCE	48.45%	22.93%	12.30%	4.15%	-2.15%	-3.40%	-2.13%	0.25%
SME ($\lambda = 10$)	**56.70%**	49.41%	42.45%	29.61%	16.29%	5.43%	0.00%	17.03%
SME ($\lambda = 50$)	54.64%	**59.70%**	56.02%	44.14%	27.22%	12.30%	2.71%	27.09%
SME ($\lambda = 100$)	52.58%	59.53%	**58.89%**	**47.74%**	29.29%	13.39%	4.10%	28.99%
SME ($\lambda = 200$)	47.42%	58.68%	58.29%	47.62%	**29.89%**	13.97%	5.04%	**29.34%**
SME ($\lambda = 300$)	44.33%	56.32%	56.89%	46.14%	29.21%	**14.20%**	5.25%	28.79%
SME ($\lambda = 400$)	43.30%	55.14%	56.48%	46.55%	28.85%	13.98%	**5.37%**	28.59%

Table 3: Detailed test accuracies for MLE, MCE, and SME using multi-condition training data.

	clean	20db	15db	10db	5db	0db	-5db	Avg.
MLE	98.52	97.35	96.29	93.79	85.52	59.00	24.50	86.39
MCE	98.79	**98.23**	97.37	95.23	86.40	60.37	24.69	87.52
SME ($\lambda = 10$)	**98.95**	98.20	**97.41**	**95.25**	**87.25**	**61.16**	**25.33**	**87.86**