A Study on Hidden Markov Model's Generalization Capability for Speech Recognition

Xiong Xiao [#1], Jinyu Li [*2], Eng Siong Chng [#3], Haizhou Li [&4], Chin-Hui Lee [♯5]

[#] *School of Computer Engineering, Nanyang Technological University, Singapore 639798*
[1] xiao0007@ntu.edu.sg [3] aseschng@ntu.edu.sg

[*] *Microsoft Corporation, Redmond, WA 98052, USA*
[2] jinyli@microsoft.com

[&] *Institute for Infocomm Research, Singapore 138632*
[4] hli@i2r.a-star.edu.sg

[♯] *School of Electrical and Computer Engineering, Georgia Institute of Technology, GA 30332, USA*
[5] chl@ece.gatech.edu

***Abstract*—From statistical learning theory, the generalization capability of a model is the ability to generalize well on unseen test data which follow the same distribution as the training data. This paper investigates how generalization capability can also improve robustness when testing and training data are from different distributions in the context of speech recognition. Two discriminative training (DT) methods are used to train the hidden Markov model (HMM) for better generalization capability, namely the minimum classification error (MCE) and the soft-margin estimation (SME) methods. Results on Aurora-2 task show that both SME and MCE are effective in improving one of the measures of acoustic model's generalization capability, i.e. the margin of the model, with SME be moderately more effective. In addition, the better generalization capability translates into better robustness of speech recognition performance, even when there is significant mismatch between the training and testing data. We also applied the mean and variance normalization (MVN) to preprocess the data to reduce the training-testing mismatch. After MVN, MCE and SME perform even better as the generalization capability now is more closely related to robustness. The best performance on Aurora-2 is obtained from SME and about 28% relative error rate reduction is achieved over the MVN baseline system. Finally, we also use SME to demonstrate the potential of better generalization capability in improving robustness in more realistic noisy task using the Aurora-3 task, and significant improvements are obtained.**

***Index Terms*—model generalization, robustness, soft margin estimation, minimum classification error, Aurora task**

I. Introduction

Speech recognition performance degrades significantly when there is mismatch between the statistics of training and testing speech due to noise distortions [1]. Traditional feature compensation [2–4] and model adaptation [5–7] methods improve the robustness of speech recognition by reducing the mismatch between training and testing conditions. Although these methods are shown to be effective, robustness of speech recognition remains as an unsolved problem.

Generalization capability of a model is a good indicator of how well the model will perform on unseen test data. From statistical learning theory [8], a big margin usually result in better generalization capability, where the margin refers to some measure of separation between competing classes. In this study, we explore the generalization capability of acoustic model to improve speech recognition robustness. A major difference between generalization capability and robustness is that generalization capability refers to model's ability to perform well on unseen but similar data as the training data (i.e. training and testing data follow the same distribution), while robustness refers to whether the model is able to perform well on unseen and mismatched testing data (i.e. training and testing data follow different distributions). Due to this difference, it is not guaranteed that good generalization capability will always results in better robustness.

Discriminative training (DT) methods are used to improve the generalization capability of acoustic model in this paper. DT methods are used as an alternative approach of the maximum likelihood (ML) method to train the hidden Markov model (HMM) based acoustic model. Generally speaking, they estimate the model parameters to reduce the empirical error (i.e. training error). Popular DT methods include minimum classification error estimation (MCE) [9, 10], maximum mutual information estimation (MMI) [11, 12], and minimum phone/word error estimation (MPE/MWE) [13]. Recently, margin-based training methods have also been applied to train acoustic models to improve generalization more explicitly, e.g. large margin hidden Markov model (LMHMM) [14, 15], large margin estimation (LME) [16], and soft-margin estimation (SME) [17]. These DT methods have also been applied to improve speech recognition robustness and shown to be effective to different extents [18–21]. In [21], model robustness is shown to be correlated to margin size, an important indicator of model's generalization capability.

In this paper, we will continue the study in [21] and examine the relationship between generalization capability of model and robustness in more details. Two discriminative training method are under our study, i.e. MCE from traditional DT methods and SME from margin-based methods. MCE focuses on reducing empirical error, while SME also considers improving the margin. We will show how these two methods improve the margin of acoustic model and how a larger margin produces better robustness empirically.

Furthermore, we address the assumption of statistical learning theory that the training and testing data are from the same distribution [8]. As this assumption is not true in noisy speech recognition, improving generalization capability not necessarily results in better robustness especially when the mismatch is big. A simple and effective remedy to this problem is to apply mean and variance normalization (MVN) [3] on speech features to reduce the mismatch before model training. We will examine how MVN will interact with MCE and SME experimentally.

In addition, we evaluate SME and MVN+SME on Aurora-3 task [22] to demonstrate how effective good generalization could be in improving robustness for more realistic tasks. Note that the noisy data in Aurora-2 task are artificially synthesized by adding recorded noises to clean speech, and the noisy data in Aurora-3 task are recorded in real noisy car environments.

The paper is organized as follows. In Section II, a brief review of SME and MCE is provided. In Section III, we present a study of how effective SME and MCE are able to improve margin and generalization. In Section IV, speech recognition results and discussions are presented. Finally, we conclude in Section V.

II. Brief Review and Comparison of SME and MCE

In SME [17, 21], the parameters of the acoustic model are estimated by minimizing the following loss function which is an approximated expected risk:

$$L^{\text{SME}}(\rho,\Lambda) = \frac{\lambda}{\rho} + R_{\text{emp}}(\rho,\Lambda) \quad (1)$$

$$= \frac{\lambda}{\rho} + \frac{1}{N}\sum_{i=1}^{N} l^{\text{SME}}(O_i,\rho,\Lambda) \quad (2)$$

$$= \frac{\lambda}{\rho} + \frac{1}{N}\sum_{i=1}^{N} \frac{\rho - d^{\text{SME}}(O_i,\Lambda)}{1+e^{-\gamma(\rho-d^{\text{SME}}(O_i,\Lambda))}} \quad (3)$$

where Λ is the set of acoustic model parameters, ρ is the soft margin, $\frac{\lambda}{\rho}$ addresses the generalization risk, and $R_{\text{emp}}(\rho,\Lambda)$ is the empirical risk (error from training data). From (2), the empirical risk is the average loss of N training utterances $O_i, i = 1, ..., N$. The separation measure $d^{\text{SME}}(O_i,\Lambda)$ represents how well the correct model is separated from its competing models regarding O_i. If the separation measure is not large enough, i.e. smaller than ρ, a loss is generated that equals to $\rho - d^{\text{SME}}(O_i,\Lambda)$. In (3), the loss of a single utterance is smoothed by a sigmoid function such that the loss function is easier to be optimized, and γ is used to control the slope of the sigmoid. The λ is used to control the relative weights of generalization risk and empirical risk. With a large λ, the training process will focus on reducing the generalization risk and the margin will be large, and vice versa.

The separation measure used in SME is the frame-normalized log likelihood ratio (LLR) [17]:

$$d^{\text{SME}}(O_i,\Lambda) = \frac{1}{n_i}\sum_{j\in F_i}\log\left[\frac{P_\Lambda(O_{ij}|S_i)}{P_\Lambda(O_{ij}|\hat{S}_i)}\right] \quad (4)$$

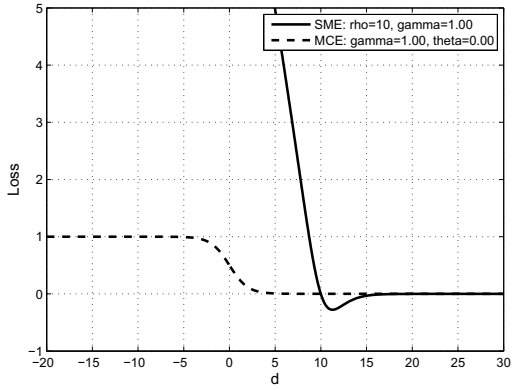

Fig. 1. Loss from a single utterance as a function of d.

where S_i and $\hat{S}_i$ represents the correct and most competing transcriptions of O_i, respectively, F_i is the set of frames in O_i that have different state identities in S_i and $\hat{S}_i$, and n_i is the number of frames in F_i.

In MCE, the loss function is defined as:

$$L^{\text{MCE}}(\rho,\Lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+e^{-\gamma d^{\text{MCE}}(O_i,\Lambda)+\theta}} \quad (5)$$

which is the average approximated classification error of training utterances. Unlike SME, there is no approximation of generalization risk in the MCE loss function. Therefore, MCE only reduces empirical risk. In this study, the following separation measure is used in MCE:

$$d^{\text{MCE}}(O_i,\Lambda) = -\log P_\Lambda(O_i|S_i) + \log\left[\frac{1}{M-1}\sum_{j,j\neq i} P_\Lambda(O_i|S_j)^\eta\right]^{1/\eta} \quad (6)$$

where M is the number of state-level alignments considered in the training, including the correct one. Note that $d^{\text{MCE}}(O_i,\Lambda)$ is not normalized by the number of confusing frames n_i while $d^{\text{SME}}(O_i,\Lambda)$ is normalized in (4).

To briefly compare SME and MCE, we plot their losses from a single utterance against the separation measure in Fig. 1. Note that the parameters here are for illustration purpose only. From the figure, MCE approximates the string classification error: when $d^{\text{MCE}}(O_i,\Lambda) > 0$, the utterance is correctly classified, and the loss is 0, and vice versa. There is a smooth transition around $d^{\text{MCE}}(O_i,\Lambda) = 0$ due to the sigmoid function. The smoothing makes the surface of the loss function continuous for easier optimization. On the other hand, the loss of SME approximates a straight line when $d^{\text{SME}}(O_i,\Lambda) < \rho$, and zero when $d^{\text{SME}}(O_i,\Lambda) > \rho$. There is some imperfect transition around $d^{\text{SME}}(O_i,\Lambda) = \rho$, which does not affect the performance of SME significantly in our study.

In our study, the loss functions of both MCE and SME are minimized by using generalized gradient descent (GPD) [17, 21], where the first order partial differentiation of the loss

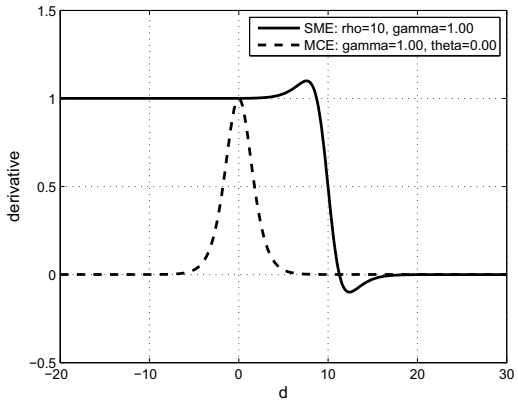

Fig. 2. Weight of a single utterance as a function of d.

function w.r.t. model parameters are used

$$\frac{\partial L}{\partial \Lambda} = \frac{\partial L}{\partial d}\frac{\partial d}{\partial \Lambda} \tag{7}$$

where L can be either $L^{\text{MCE}}(\rho, \Lambda)$ or $L^{\text{SME}}(\rho, \Lambda)$, d can be either $d^{\text{MCE}}(O_i, \Lambda)$ or $d^{\text{SME}}(O_i, \Lambda)$. We focus our comparison here on $\frac{\partial L}{\partial d}$, which will explain the major difference between MCE and SME. The gradient of the acoustic model parameters from one utterance can be represented as follows:

$$\frac{\partial l_i^{\text{SME}}}{\partial d} = \frac{-1}{1+e^{-\gamma(\rho-d)}}\left[1+\frac{\gamma(\rho-d)e^{-\gamma(\rho-d)}}{1+e^{-\gamma(\rho-d)}}\right] \tag{8}$$

$$\frac{\partial l_i^{\text{MCE}}}{\partial d} = \frac{-\gamma e^{-\theta+d\gamma}}{(e^{d\gamma}+e^{-\theta})^2} \tag{9}$$

where l_i^{SME} is the simplified notations for $l^{\text{SME}}(O_i, \rho, \Lambda)$ and l_i^{MCE} is the loss of MCE due to utterance O_i. The negative values of $\partial l_i^{\text{SME}}/\partial d$ and $\partial l_i^{\text{MCE}}/\partial d$ can be seen as the weights of utterance O_i in the estimation process of SME and MCE respectively. These two variables are plotted against d in Fig. 2. Note that the maximum weight for MCE is normalized to 1 for better comparison. From the figure, SME approximates a step function, where the weight of an utterance is zero if $d > \rho$ and 1 otherwise. A larger ρ will increase the separation measures more aggressively and produce larger margin in the resulting model. In our study, we find that the imperfect transitions around ρ do not affect the performance significantly. Unlike SME, MCE only assigns large weights around θ/γ. As θ is usually set to zero, MCE uses mainly utterances around the decision boundary. The sigmoid slope is controlled by γ and is very important for the generalization capability of MCE-trained model. A smaller γ will make the plot of MCE in Fig. 2 fatter and hence utterances far from the decision boundary will be active during training. As a result, a smaller γ produces larger margin and better generalization capability.

III. Effect of SME and MCE on Separation Measures

Let's first look at how SME and MCE improves the separation measure. Note that the separation measure used here is computed using (4), even for MCE, for a more consistent comparison. Both SME and MCE are implemented using GPD. N-best competing transcriptions (N=2) are used as the source of confusion patterns. The features are processed by MVN [3] in an utterance-by-utterance fashion.

Fig. 3(a) shows the histograms of separation measures of 8440 clean training utterances defined by Aurora-2 task. From the figure, it is observed that the histograms obtained with both SME model and MCE model are shifted right significantly compared to that obtained with ML model, while the improvement of SME is much bigger than that of MCE. This may indicate a better generalization capability of the SME and MCE models than the ML model. Furthermore, SME and MCE also reduce empirical error significantly, as demonstrated by the reduced histogram on the left hand side of the xy-plane (x=0).

In Fig. 3(b), the same study is carried out on the clean test data. There are totally 10,010 test utterances in the clean test set, the same as the following 10dB and -5dB test sets. Compared to ML, SME significantly increases the separation measures of testing data as it has a larger margin than ML as shown in Fig. 3(a). Similarly, MCE also improves the separation measures of testing data over ML, but less significant than SME.

Fig. 3(c) shows the histograms obtained from 10dB test data. It is observed that the effect of the two DT methods becomes less significant in 10dB test set than in clean test set. One reason for this observation is that, as the mismatch becomes larger, the margin becomes less effective in covering the mismatch. Another reason is that the confusion pattern of noisy testing data may be different from that of clean training data. Hence, what SME and MCE learn from training data becomes less relevant on noisy testing data. In addition, it is observed that the improvement of SME is still significantly better than that of MCE, this should be due to the larger margin of SME than MCE on the training data shown in Fig. 3(a).

In Fig. 3(d), the histograms of -5dB test sets are shown. As the SNR level is extremely low, noise is more dominant than speech, and both SME and MCE actually decrease the mean of the histogram. This shows that when the mismatch is too big, the generalization capability will fail to improve robustness as it is meant to be applied on matched testing data. However, the area under the histogram on the right of x=0 is increased by SME and MCE, indicating a better string accuracy. The reason may be that SME and MCE is able to improve separation measures for those utterances in relatively good condition, while it degrades separation measures for bad utterances.

In summary, both SME and MCE are able to improve the margin of the model on training data significantly. The improved margin indicates a better generalization of the model, which results in better robustness on the test data, especially for less mismatched cases. In addition, SME is shown to be more effective than MCE. This should be due to the fact that SME is designed to explicitly increase the margin, while MCE focuses mainly on reducing empirical error.

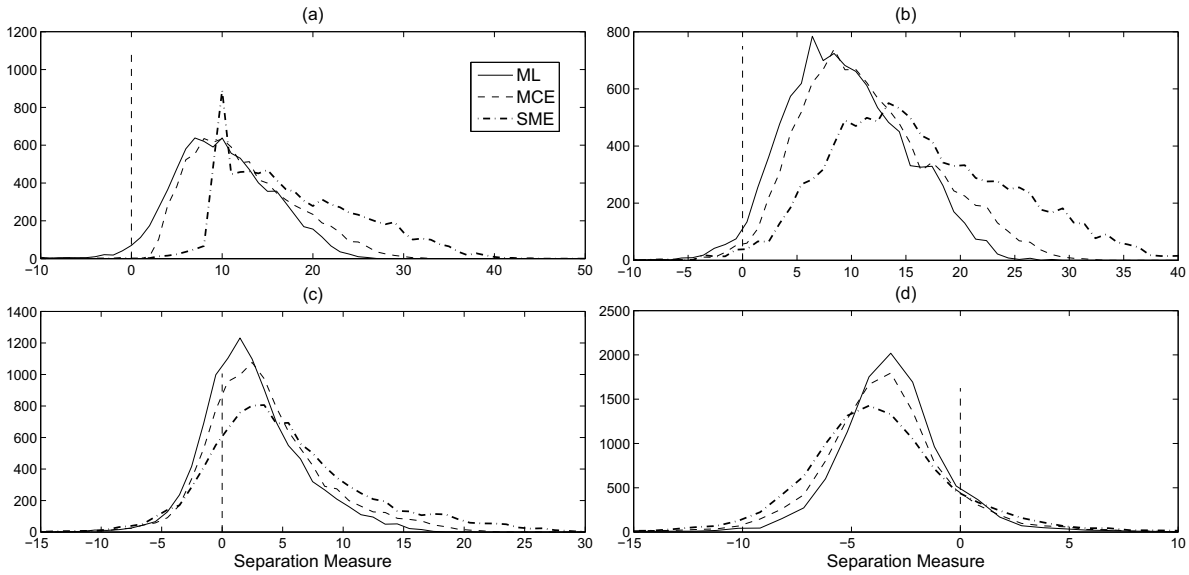

Fig. 3. Effect of SME and MCE on the separation measures of speech. For each of the 4 scenarios from Aurora-2 task under study, separation measures of utterances belonging to the scenario are computed using (4) with ML-, SME-, and MCE-trained acoustic models and represented as histograms. The y-axis (i.e. x=0) is also plotted for analysis. The separation measures are collected from: (a) clean training data; (b) clean testing data; (c) 10dB test data; (d) -5dB test data.

IV. Speech Recognition Experiments

A. System Description

The performance of SME and MCE is evaluated on Aurora-2 [23] and Aurora-3 [22] tasks. Aurora-2 is a noisy English connected digit task, where the noisy speech data are artificially generated by adding recorded noises to clean speech. In Aurora-3 task, there are 5 noisy connected digit sub-tasks, each for one European language. Furthermore, the data in Aurora-3 task are all recorded in real noisy car environments. Therefore, Aurora-2 and Aurora-3 tasks are complementary.

The acoustic models use standard "simple back-end" configurations but without short pause model. Each digit is modeled by a 16-state HMM with 3 Gaussian mixtures per state. Mel-frequency cepstral coefficients (MFCC) are used as features and extracted using the WI007 feature extraction program provided by Aurora-2. There are 39 raw features, including 13 static features and their first and second order differential features. Cepstral energy C0 is used instead of log energy (This is slightly different from the system in [21]).

The value of λ in (1) is set to be 5 for all following experiments as SME performance is not very sensitive to λ if λ is within a proper range. Readers are referred to [21] for a more detailed examination of λ's effects on Aurora-2 task. The γ used for SME in (3) is set to 2. For MCE, γ affects the resulting model's generalization significantly. We empirically set γ to 0.05 for clean condition training and 0.1 for multi-condition training of Aurora-2 task, respectively. Note that γ used in SME is very different from γ used in MCE as $d^{\text{MCE}}(O_i, \Lambda)$ is not normalized and $d^{\text{SME}}(O_i, \Lambda)$ is normalized.

B. Performance with Raw MFCC Features on Aurora-2

The performance of SME and MCE on Aurora-2 task is shown in Table I. Note that the R.R. columns represent the relative reduction of word error rate (WER) achieved by SME over ML baseline, i.e. R.R. $= \frac{\text{WER}_{\text{ML}} - \text{WER}_{\text{SME}}}{\text{WER}_{\text{ML}}} \times 100$. From the table, both SME and MCE improve recognition accuracy significantly over ML for both clean and multi-condition training schemes, with SME delivers moderately better results than MCE. This observation shows the effectiveness of generalization capability for improving robustness of speech recognition. Different improvement patterns are observed on the two training schemes. In clean condition training, SME and MCE performs better at high SNR levels (15dB and above) than at low SNR levels (5dB and below). In multi-condition training, as the training data include noisy data down to 5dB, more even improvements are observed at all SNR levels. This shows that improving model's generalization capability is able to cover moderate mismatch but less effective in covering large mismatch due to the limited margin for handling mismatch as shown in Figure 3(a). As speech recognition is a multi-class problem, there is a limit on the margin we can obtain.

We also examine the performance of SME and MCE for the 3 test sets in multi-condition training scheme in Table II. The 4 types of noises in test set A are observed during training while the 4 types of noises in test set B are not. The test set C is corrupted by both additive noise and convolutive channel distortion. One of the two noises in test set C is observed during training and the other is not. Generally speaking, test set A represents the most matched scenario, and test set B represents the most mismatched scenario. From Table II, it is observed that SME and MCE general produce the highest improvement for test set A and the least improvement for

TABLE I
PERFORMANCE OF SME AND MCE WITH RAW MFCC FEATURES ON AURORA-2 TASK. WORD ACCURACIES OF BOTH CLEAN AND MULTI-CONDITION TRAINING SCHEMES ARE SHOWN AT DIFFEREN SNR LEVELS. ML REPRESENTS THE MAXIMUM LIKELIHOOD BASELINE. *R.R.* REFERS TO THE RELATIVE REDUCTION OF WORD ERROR RATE ACHIEVED BY SME OVER THE CORRESPONDING ML BASELINE RESULTS.

SNR	Clean Condition				Multi-Condition			
	ML	MCE	SME	*R.R.*	ML	MCE	SME	*R.R.*
Clean	99.04	99.61	99.64	*62.04*	98.60	99.08	99.13	*37.89*
20dB	94.36	97.40	97.67	*58.71*	97.66	98.44	98.67	*43.11*
15dB	85.58	92.27	92.95	*51.11*	96.69	97.81	98.05	*41.17*
10dB	66.82	75.23	76.32	*28.64*	94.38	95.66	96.38	*35.66*
5dB	39.20	45.86	47.32	*13.35*	86.77	89.31	90.18	*25.81*
0dB	17.14	21.56	22.93	*6.99*	59.46	65.99	66.51	*17.39*
-5dB	9.78	10.98	11.70	*2.12*	24.27	26.49	26.65	*3.14*
0-20dB	60.62	66.47	67.44	***17.31***	86.99	89.44	89.96	***22.82***

TABLE II
PERFORMANCE OF SME AND MCE WITH RAW FEATURES ON THE 3 TEST SETS OF AURORA-2 TASK USING MULTI-CONDITION TRAINING. ACC. REFERS TO AVERAGE WORD ACCURACY AND R.R. IS THE RELATIVE WER REDUCTION ACHIEVED OVER THE CORRESPONDING ML BASELINES.

Test Set	ML	MCE		SME	
	Acc.	Acc.	R.R.	Acc.	R.R.
A	87.60	91.68	***32.93***	91.93	***34.90***
B	87.86	88.87	***8.34***	89.44	***13.01***
C	84.04	86.11	***12.94***	87.07	***18.96***
Average	86.99	89.44	*18.85*	89.96	*22.82*

test set B. This is reasonable given their different levels of mismatches with the training data. It is also observed that SME has more advantage over MCE in test set B and C than in test set A. This shows that for more mismatched scenarios, SME may have a larger advantage over MCE as it is more aggressive in increasing margin.

C. Interaction with MVN

The effect of generalization capability for improving robustness is limited when there is large mismatch between the training and testing data. In this section, we will reduce the mismatch and examine its effect. MVN [3] is used to process both the training and testing features before model training and testing. Each of the 39 MFCC features are processed by utterance-based MVN individually.

The performance of the combined system is shown in Table III. Comparing Table III with Table I, we can observe that SME produces even better improvement when MVN is used in terms of relative error reduction (R.R.), especially at low SNR levels. This is because when the mismatch is reduced by MVN, it is easier for SME to cover the mismatch. Similar trend is also observed for MCE. It may be a good strategy to combine feature domain robustness techniques with SME and MCE for better robustness.

Similar to Table II, we also examine the performance of SME and MCE on the three test sets for multi-condition training scheme in Table IV. Compare the results of these two tables, it is observed that the performance gap between test set A and test set B and C becomes smaller after MVN processing. This is because after the MVN processing, the feature distortion is reduced and training set can better

TABLE III
PERFORMANCE OF SME AND MCE WITH MVN-PROCESSED FEATURES ON AURORA-2 TASK.

SNR	Clean Condition				Multi-Condition			
	ML	MCE	SME	*R.R.*	ML	MCE	SME	*R.R.*
Clean	99.16	99.59	99.68	*61.86*	98.23	99.08	99.20	*54.80*
20dB	97.42	98.34	98.51	*42.19*	98.53	99.15	99.28	*51.19*
15dB	95.17	96.67	96.85	*34.76*	97.70	98.77	98.93	*53.71*
10dB	89.34	92.06	93.09	*35.16*	96.09	97.63	97.92	*46.67*
5dB	74.48	80.15	82.93	*33.12*	90.71	93.66	94.02	*35.63*
0dB	45.21	54.06	58.67	*24.57*	74.26	78.98	79.28	*19.49*
-5dB	17.81	22.76	24.90	*8.63*	40.87	44.66	45.43	*7.70*
0-20dB	80.33	84.25	86.01	***28.89***	91.46	93.64	93.89	***28.42***

TABLE IV
PERFORMANCE OF SME AND MCE WITH MVN-PROCESSED FEATURES ON THE THREE TEST SETS OF THE AURORA-2 TASK USING MULTI-CONDITION TRAINING.

Test Set	ML	MCE		SME	
	Acc.	Acc.	R.R.	Acc.	R.R.
A	91.44	93.92	***28.97***	94.10	***31.07***
B	91.64	93.42	***21.29***	93.81	***25.96***
C	91.13	93.51	*26.83*	93.61	*27.96*
Average	91.46	93.64	*25.53*	93.89	*28.42*

represent test set B and C. Furthermore, it is also observed in Table IV that the improvement of SME over MCE is larger for test set B than for test set A. This is similar to the observation in Table II and further shows the advantage of SME.

D. Performance on Aurora-3

We also evaluate SME on Aurora-3 task, in which the data were recorded in real noisy environments. As we have shown that SME is a more direct way of improving model's generalization capability and performs better than MCE on Aurora-2 task, we will only show results of SME on Aurora-3 task in this section for brevity.

The performance of SME with raw MFCC features is shown in Table V. From the results, we have two observations. First, SME improves performance for all cases except for the high-mismatch (HM) of German. This suggests that better generalization capability is also able to improve robustness for realistic tasks. Second, SME produces higher improvement when the mismatch is relatively low. The improvements for well-match (WM) sub-tasks are always the highest, followed by medium-mismatch (MM), and improvements are usually the lowest for HM. The mismatch in HM may be beyond the generalization capability of the SME-trained acoustic model to tolerate. Similar results are observed on Aurora-2 (Table I), where performance at very low SNR levels is usually less improved due to the high level of mismatch.

The performance of SME with MVN-processed MFCC features is shown in Table VI. Although the improvements are quite different for different sub-tasks due to their different characteristics, on average, SME with MVN delivers better performance improvement than SME alone. This further shows the complementary effects of SME and MVN.

V. CONCLUSION AND FUTURE WORK

In this paper, we have shown that by improving the margin and generalization capability of the acoustic model using

TABLE V
PERFORMANCE OF SME WITH MFCC ON AURORA-3 TASK. THE THREE TRAINING SCHEMES ARE: WELL-MATCHED (WM), MEDIUM-MISMATCH (MM) AND HIGH-MISMATCH (HM). IN AVERAGED RESULTS, THE WEIGHTS OF WM, MM AND HM ARE 40%, 35% AND 25%, RESPECTIVELY.

Scheme	Finnish			Spanish			German			Danish			Italian		
	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*
WM	92.00	96.97	*62.13*	86.08	94.69	*61.85*	90.62	92.59	*21.00*	77.92	89.24	*51.28*	94.70	97.02	*43.77*
MM	69.36	78.39	*29.47*	73.28	84.53	*42.10*	79.28	80.97	*8.16*	53.11	64.41	*24.09*	85.30	86.38	*7.35*
HM	42.61	56.47	*24.15*	41.29	54.05	*21.73*	72.66	72.66	*0.00*	38.01	43.14	*8.28*	40.58	45.62	*8.48*
Avg.	71.73	80.34	***30.47***	70.40	80.97	***35.72***	82.16	83.54	***7.73***	59.26	69.02	***23.97***	77.88	80.45	***11.60***

TABLE VI
PERFORMANCE OF SME WITH MVN-PROCESSED MFCC FEATURES ON AURORA-3 TASK.

Scheme	Finnish			Spanish			German			Danish			Italian		
	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*	ML	SME	*R.R.*
WM	89.24	97.82	*79.74*	93.16	96.35	*46.64*	93.01	94.27	*18.03*	85.12	91.82	*45.03*	94.59	97.79	*59.15*
MM	76.68	89.12	*53.34*	86.55	89.28	*20.30*	84.63	85.21	*3.77*	62.71	71.47	*23.49*	82.26	90.81	*48.20*
HM	79.65	82.90	*15.97*	81.65	83.31	*9.05*	86.63	86.63	*0.00*	62.38	72.49	*26.88*	81.02	83.99	*15.65*
Avg.	82.45	91.05	***48.98***	87.97	90.62	***22.00***	88.48	89.19	***6.14***	71.59	79.87	***29.12***	86.88	91.90	***38.23***

SME and MCE, the robustness of speech recognition can be improved significantly. Our results also showed that improving generalization capability is complementary to traditional feature normalization method MVN. Our results on Aurora-2 and Aurora-3 tasks are very attractive given the fact that there is no online adaptation during testing.

There are several issues needed to be investigated in the future. Although margin/generalization is shown to be correlated with robustness, big improvement of margin achieved by SME does not deliver proportionally big improvement in robustness when we compare SME and MCE (see Fig. 3 and Table III). One reason may be that the current separation measure only considers the most competing transcription and may not be able to represent the true generalization capability of the model. Besides, both Aurora-2 and Aurora-3 are connected digit tasks. It will be interesting to investigate whether the observations in this paper also applies to more complex tasks.

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[4] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 845–854, 2006.

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

[6] J. L. Gauvain and C. H. Lee, "Maximum *a posterirori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[7] Y. Gong, "A method of joint compensation of additive and convolutive distortions for speaker-independent speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 975–983, 2005.

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

[9] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, no. 12, pp. 3043–3054, Dec 1992.

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

[11] L. R. Bahi, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Proc. ICASSP '86*, Tokyo, Japan, Apr. 1986, pp. 49–52.

[12] V. Valtchev, J. J. Odell1, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303–314, 1997.

[13] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 2003.

[14] F. Sha and L. Saul, "Large margin gaussian mixture modeling for phonetic classification and recognition," in *Proc. ICASSP '06*, vol. 1, May 2006, pp. I–I.

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

[16] H. Jiang, X. Li, and C. Liu, "Large margin hidden markov models for speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1584–1595, 2006.

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

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

[19] B.-W. Mak, Y.-C. Tam, and P. Li, "Discriminative auditory-based features for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 1, pp. 27–36, Jan. 2004.

[20] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proc. InterSpeech '05*, Lisbon, Portugal, Sep. 2005, pp. 989–992.

[21] J. Li and C.-H. Lee, "On a generalization of margin-based discriminative training to robust speech recognition," in *Proc. InterSpeech '08*, Brisbane, Australia, Sep. 2008.

[22] *Baseline results for subsect of SpeechDat-Car Finnish database for ETSI STQ WI008 advance front end evaluation*, Aurora document no. AU/255/00, Nokia, Jan 2000.

[23] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recogntion systems under noisy conditions," in *Proc. ICSLP '00*, vol. 4, Beijing, China, Oct. 2000, pp. 29–32.