A Study on the Generalization Capability of Acoustic Models for Robust Speech Recognition

Xiong Xiao*, *Student Member, IEEE,* Jinyu Li, *Member, IEEE,* Eng Siong Chng, *Senior Member, IEEE,* Haizhou Li, *Senior Member, IEEE,* Chin-Hui Lee, *Fellow, IEEE,*

***Abstract*—In statistical learning theory, good generalization capability refers to small performance degradation when the model is evaluated on unseen testing data that are drawn from the same distribution as the training data, i.e. on matched training-testing case. Recently, soft-margin estimation (SME) method was proposed to improve acoustic model's generalization capability for clean speech recognition and achieved success. In this paper, we study the generalization capability of acoustic model for robust speech recognition, where the training and testing data follow different distributions (i.e. mismatched training-testing case). From our analysis of noise effect on the log likelihood values of noisy speech features, although mismatch exists between testing and training data, it is still possible to achieve better robustness by improving the acoustic model's generalization capability using SME. This is confirmed by our experimental study on Aurora-2 and Aurora-3 tasks, where SME improves recognition performance significantly for both matched and low/medium mismatched testing cases. However, the improvement in severely mismatched cases is relatively small. To alleviate the violation of SME assumption about the same distribution for training and testing data, we apply mean and variance normalization (MVN) to process speech features prior to model training. Experimental study shows that when training-testing mismatch is reduced, SME delivers better performance improvement. We expect SME to improve the robustness of speech recognition further when it is combined with other robustness methods. Although this study is on noisy speech recognition tasks, the method and discovery in this paper have no assumption on the type of distortion, and can be extended to deal with different types of distortions in other machine learning applications.**

EDICS Category: SPE-ROBU; SPE-RECO

I. Introduction

Speech recognition performance degrades significantly when speech signals are corrupted by noises [1]. The noise distortion usually causes a difference between the statistics of training and testing speech features. Typically, the acoustic model of a speech recognition system is trained from clean data using the maximum likelihood (ML) criterion. Hence, the decision boundary of the model fits well to the distribution of the clean training data. However, when noisy testing data with different distribution are tested, the decision boundary may fail and recognition performance will degrade.

To improve the robustness of speech recognition against noise distortions, many methods have been proposed to reduce the mismatch between clean-trained model and noisy testing data. These methods can be grouped into two classes, i.e. feature compensation methods and model adaptation methods. The feature compensation methods aim to make the features from different environmental conditions more consistent while preserving the features' discriminative power. Such methods include various speech parameter estimators [2–8]; feature normalization methods: cepstral mean normalization (CMN) [9], mean and variance normalization (MVN) [10], histogram equalization (HEQ) [11–14]; temporal filters: RASTA filter [15], MVA processing [16] and temporal structure normalization filter (TSN) [17, 18]; etc. In contrast, the model adaptation methods reduce the mismatch by making the acoustic model better fit the noisy testing data such that the adapted decision boundary is more accurate for the noisy testing data. Typically, the parameters of the acoustic model are adapted based on observed noise data. Model adaptation methods include: maximum likelihood linear regression (MLLR) [19] adaptation, maximum *a posteriori* (MAP) adaptation [20], parallel model composition (PMC) [21], ensemble modeling [22, 23], and joint compensation of additive and convolutive distortions (JAC) [24, 25], etc.

Although the feature compensation and model adaptation methods are quite effective, reducing mismatch is not the only way to improve the robustness of speech recognition. In this paper, we follow another direction to improve robustness, i.e. improving the generalization capability of the acoustic model. Instead of pursuing a good fit of the acoustic model to the training data as in the ML estimation, we estimate the model parameters to make the model more generalizable to unseen testing data.

According to statistical learning theory [26], the generalization capability of model can be improved by increasing the margin of the model. The margin is the desired minimum distance between any training sample to the decision boundary of the model in a separable classification case. During the model training, model parameters are estimated such that all or most training samples are outside of the margin. As a result, a buffer zone is created around the decision boundary, and the model becomes more generalizable. A large margin corresponds to a more general model. Recently, the soft-margin estimation (SME) method [27] was proposed to maximize the margin for

This work was performed during the first author's internship in School of Electrical and Computer Engineering, Georgia Institute of Technology, USA in 2008.

Xiong Xiao, Eng Sion Chng and Haizhou Li are with School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue 50, Singapore 639798. Jinyu Li is with Microsoft Corporation, One Microsoft way, Redmond, WA, 98052, USA. Haizhou Li is also with the Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613. Chin-Hui Lee is with School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA. (email: xiao0007@ntu.edu.sg, jinyli@microsoft.com, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg, chl@ece.gatech.edu.)

speech recognition problems and shown to perform well in clean speech recognition.

In statistical learning theory, the generalization of model refers to generalizing to testing data drawn from the same distribution as the training data. In noisy speech recognition problem to be studied in this paper, the training and testing are from different distributions. Therefore, the assumption of same distribution for both training and testing features required by SME is violated in noisy speech recognition tasks. However, we will show that increasing the margin of the acoustic model is still desirable in mismatched training-testing cases and our experimental results will verify the effectiveness of this approach. In [28], we have conducted an initial study of SME for Aurora-2 task [29] and achieved promising results. In this paper, we conduct a more complete study of the approach of improving model generalization for better robustness against noise corruption. Furthermore, we will also study the combination of SME with mean and variance normalization (MVN). As MVN is able to reduce the mismatch between training and testing data, if SME operates on MVN-processed features, the assumption of SME about the same distribution for training and testing data will be less violated. We expect SME to perform better when combined with MVN, or other feature domain methods.

This paper is organized as follows. In section II, we investigate the noise effect in log likelihood domain and decision-making in speech recognition. This investigation provides insight and motivation to the use of margin-based model training approach. In section III, we discuss the method of margin maximization for more robustness model and describe the SME method. In section IV, we present our experimental results and discussions. Finally, we conclude in section V.

II. Noise Effect on Log Likelihoods

When a speech signal is corrupted by noise, the speech features extracted from the speech signal are also distorted. The likelihood values of the distorted features evaluated on different classes of clean-trained acoustic model will be different from those of clean features. Therefore, the classification decision based on the changed likelihood values will not be optimal. In this section, we will analyze the noise effect in the log likelihood domain, and show the necessity to reduce noise effect on log likelihoods.

A. A Two-Class Example

Speech recognition is a multi-class sequential pattern recognition problem. The temporal dynamics of speech and the use of hidden Markov models (HMM) make the direct analysis of noise effect on log likelihoods very difficult. In this section, we will first use a two-class, single feature vector based pattern classification problem as an example to demonstrate noise effect, and then examine noise effect in real speech recognition experimentally.

Let there be two classes "A" and "B" as shown in Fig. 1(a). We assume there are N training samples, each has two elements $\{X_i, C_i\}$, where X_i is a feature vector of D dimensions, C_i is the correct class label of X_i and $C_i \in \{A, B\}$. We want

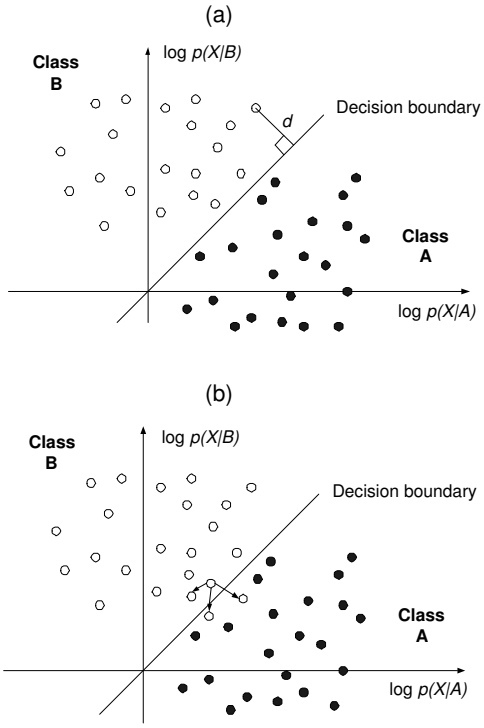

Fig. 1. Illustration of the two-class classification problem in log likelihood domain: (a) if well trained, the model is able to project the clean feature vectors to the correct side of the decision boundary; (b) when noise distortion presents, the noisy samples will deviate from the clean samples and may cross the decision boundary, thus wrongly classified.

to build a classifier that can correctly classify any unseen feature vector into one of the two classes. A common way to build a classifier for this problem is to first estimate the probability density function (p.d.f.) of feature vectors for each class and then use the maximum *a posteriori* (MAP) decision rule to classify the testing samples. The classification decision for a feature vector X_i is made as follows:

$$\begin{aligned} \hat{C}_i &= \arg \max_{j \in \{A,B\}} p(j|X_i) \\ &= \arg \max_{j \in \{A,B\}} p(X_i|j)p(j) \end{aligned} \quad (1)$$

where $p(j)$ is the *a priori* probability of class j, i.e. our prior knowledge about class j, and $p(j|X_i)$ is the *a posteriori* probability of class j after X_i is observed. In speech recognition, the *a prior* knowledge about classes, such as words, are represented by language model. As we are only interested in the noise effect in acoustic modeling, it is reasonable to ignore the language model in our analysis. We assume the two classes have equal *a priori* probability and (1) can be rewritten as follows:

$$\hat{C}_i = \arg \max_{j \in \{A,B\}} p(X_i|j) \quad (2)$$

and this is the maximum likelihood (ML) decision rule illustrated in Fig. 1(a). The decision boundary is the straight line $\log p(X_i|A) = \log p(X_i|B)$. The x-axis and y-axis represent the log likelihoods of training samples on class A and B, respectively. In the log likelihood domain, the classification decision is based on the Euclidian distances from samples to

the decision boundary. The distance between a sample of class A to the decision boundary is

$$\begin{aligned} d(X_i,\Lambda) &= \frac{\sqrt{2}}{2}[\log(p(X_i|A)) - \log(p(X_i|B))] \\ &= \frac{\sqrt{2}}{2} d^{LLR}(X_i,\Lambda) \end{aligned} \quad (3)$$

where $d^{LLR}(X_i,\Lambda)$ is the log likelihood ratio (LLR) of X_i on model $\Lambda = \{\lambda_A, \lambda_B\}$, and λ_A and λ_B denote the parameters of the p.d.f. of class A and B, respectively. If X_i is from class B, $d^{LLR}(X_i,\Lambda) = \log(p(X_i|B)) - \log(p(X_i|A))$. If $d(X_i,\Lambda) > 0$, X_i is correctly classified and vice versa. The distance serves as a measure of separation, i.e. how well a training sample is separated from the decision boundary by the model. If a training sample is far from the decision boundary, the sample is well separated by the model.

The model Λ is like a transformation, which transforms a D-dimensional feature vector into a two-dimensional vector whose elements are the coordinates of the feature vector in the log likelihood domain. Usually the transformation is trained to project the training samples into the correct side of the decision boundary. If testing features have similar probability distribution as that of the training features, they can also be projected correctly. However, this is not true if the testing features have different probability distribution, e.g. due to noise corruption.

When a speech signal is corrupted by noise, the features extracted from the signal will also be distorted. If we assume the distortion to be additive and independent from clean features in the feature domain, the noisy features can be represented as:

$$Y_i = X_i + N_i \quad (4)$$

where Y_i is the corrupted feature vector and N_i is the distortion in feature domain. The distortion N_i will cause disturbance of $\log p(Y_i|A)$ and $\log p(Y_i|B)$, i.e. the two coordinates of Y_i in the log likelihood domain. Therefore, Y_i will wander off X_i and the distance between them is governed by a probability distribution. This is illustrated in Fig. 1(b), where we randomly show three possible deviations of noisy sample from a clean sample. Although the clean sample is on the correct side of the decision boundary, its noisy versions may cross the boundary and be wrongly classified. It is reasonable to predict that the lower the signal-to-noise ratio (SNR) in the signal domain, the higher the variance of N_i in the feature domain, and possibly the larger deviation of Y_i away from X_i in the log likelihood domain. With larger deviation, the test samples are more likely to be projected into the wrong side of the decision boundary, thus wrongly classified.

B. Empirical Study of Noise Effect on Speech Recognition

Speech recognition problem is far more complex than the two-class problem. There are several major differences between the two which can be summarized as follows:

1) Speech patterns are represented by a sequence of feature vectors rather than a single feature vector. To model the temporal dynamics of speech, complicated HMM is used as the model.

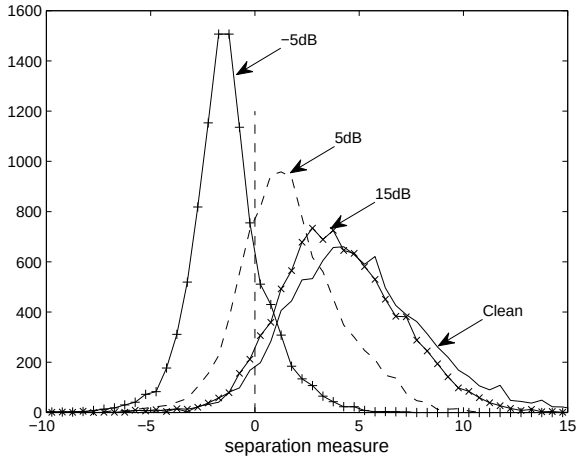

Fig. 2. Histogram of separation measures for different SNR levels. Each histogram is obtained from 10,010 separation measure instances.

2) In speech recognition, there are many classes rather than two classes. With N classes, the features are projected into N-dimensional vectors in log likelihood domain rather than two-dimensional vectors. It is hard to carry on the study unless we only consider the correct and the closest competing classes.
3) The assumption that the noise term is additive and independent from the speech is also not true in real feature extraction of speech recognition systems, such as Mel-frequency cepstral coefficients (MFCC). It is well known that in the cepstral domain, the relationship between noise and speech is highly nonlinear [30, 31].

With the above challenges among many others, it is mathematically difficult to study the noise effect for real speech recognition system. Nevertheless, the two-class example in the previous sections provides us an intuitive example of noise effect. In this section, we will empirically study how noise affects the log likelihood of features in speech recognition.

Our study of noise effect will be described now. The histogram of separation measures of training samples will be shown. Note that each sample is an utterance in speech recognition. The calculation of separation measure is described as follows. For each utterance, we find the correct state-level alignment of the utterance using correct transcription and the acoustic model trained from clean features. We also find the closest competing alignment of the utterance using the clean model. The next step is to find out the frames with confusion, i.e. the frames that have different state identities in the correct and competing alignments. The separation measure is defined as the average log likelihood ratio (LLR) of those selected frames [27]:

$$d(O_i,\Lambda) = \frac{1}{n_i} \sum_{j \in F_i} \log \left[\frac{P_\Lambda(O_{ij}|S_i)}{P_\Lambda(O_{ij}|\hat{S}_i)} \right] \quad (5)$$

where Λ is the clean acoustic model, O_{ij} is the j^{th} frame of the i^{th} utterance O_i; S_i and $\hat{S}_i$ represent the correct and the closest competing alignments of O_i, respectively; F_i is

the set of frames in O_i with confusion; and n_i is the number of frames in F_i. Note that the separation measure $d(O_i, \Lambda)$ is related to the distance between a training sample and the decision boundary [see (3)] .

We study the noise effect on separation measure using the Aurora-2 task [29]. The testing data of Aurora-2 are divided into seven groups according to SNR level, including clean testing data, 20dB to -5dB testing data with 5dB step. In each SNR level, there are 10,010 utterances, each corrupted by one of 10 types of noises. For each utterance, we obtain its separation measure as described in (5), using acoustic model trained from clean data and the ML estimation. The histogram of separation measures for different SNR levels are compared in Fig. 2. The part of histogram on the left side of the vertical line at 0 are for those wrongly classified utterances. From the figure, we observe that as SNR level decreases, the histogram of separation measures shift left and becomes sharper. This shows that in overall, the distances between test samples and the decision boundary are reduced by noise distortion and some utterances are moved to the wrong side of the decision boundary. With lower SNR level, there are higher distortion in the feature domain, and possibly larger deviation of noisy samples from clean samples in the log likelihood domain, and hence the histogram of separation measures are shifted left further.

C. Summary

Noise corruption is shown to cause noisy features to deviate from corresponding clean features randomly in the log likelihood domain. As a result, when noisy features are tested on clean trained models, or more generally, whenever there is statistical mismatch between training and testing features, recognition performance will degrade. To improve the robustness of speech recognition systems against noise distortion, it is necessary to make the recognition less sensitive to noise effect in the log likelihood domain.

III. Improving the Generalization Capability of Acoustic Model

Currently, the noise effect is reduced by either feature compensation methods or model adaptation methods. In feature compensation methods, if we can obtain an accurate estimate of the clean feature from the observed noisy features, the deviation of log likelihood will be reduced and better classification can be performed. In model adaptation methods, the model are adapted to approximate the model trained from the noisy test features. If the adapted model can represent the noisy test features well, the projection of feature vectors to log likelihood domain will also be correct and performance will be improved. Although both feature compensation and model adaptation are very important and effective ways of reducing noise effect, we are going to propose another approach for the problem. We aim at improving the generalization capability of the acoustic model, i.e. the robustness of the projection of acoustic model. In this section, we will first introduce the concept of improving the generalization capability of acoustic model, and then describe the SME method used to achieve our objective.

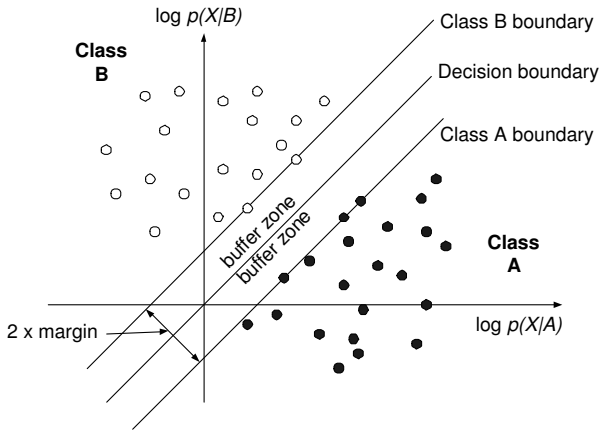

Fig. 3. Increasing margin to improve the generalization capability of the model. The objective is to adjust the model parameters to pull the training samples out of the class boundaries defined by the margin. As a result, a buffer zone will be created around the decision boundary, hence the model will be more robust against the deviations caused by noise as shown in Fig. 1(b).

A. Generalization Capability and Margin

By generalization capability, we refer to the ability of the model to generalize well to data that are not observed during training. When generalization capability of acoustic model is improved, the speech recognition system is more likely to perform well on mismatched test data.

Statistical learning theory [26] provides us some insights about improving pattern classification systems' generalization capability. In this theory, the expected risk of a system is formulated as

$$R(\Lambda) \leq R_{emp}(\Lambda) + R_{gen}(\Lambda) \tag{6}$$

where the empirical risk $R_{emp}(\Lambda)$ is the system's recognition error on training data and the generalization risk $R_{gen}(\Lambda)$ is a regularization term proportional to model complexity. Expected risk refers to the recognition error of the system on all data in the problem scope, i.e. both clean and noisy speech data in the case of noisy speech recognition. Both empirical and generalization risks are related to model complexity. For example, a more complex model is able to fit better to training data to produce lower empirical risk, however, it also leads to higher generalization risk. Minimum expected risk is obtained when a good balance between these two risks is achieved.

According to statistical learning theory, the generalization risk is bounded by a function which is proportional to model complexity. For the bound to be true, some assumptions are required, e.g. the training and testing data are generated from the same identical and independent distribution. However, in robust speech recognition problems, the assumption is not true, hence the bound does not exist. Fortunately, the lack of a bound does not prevent us from reducing the generalization risk for mismatched problems. In fact, even when the assumption is true and a bound exists, the bound is usually not very useful in practical classifier design due to difficulties in evaluating the bound. Instead, the reducing of generalization risk relies on another factor, the margin of the model.

The generalization risk can be reduced if margin is increased [26] as illustrated in Fig. 3. Margin serves as a

desired minimum distance between training samples and the decision boundary. During model training, the objective is to pull those training samples that fall within the margin away from the decision boundary. Those samples already far from the decision boundary do not contribute to model parameter estimation. After training, all or most training samples will be outside the margin, and a "buffer zone" is formed around the decision boundary with width equal to the margin in each side. With this "buffer zone", if a test sample deviates from the training samples of its correct class but the distance between the test sample and its nearest training sample is less than the margin, correct decision can still be made. If a larger margin is used during training, the "buffer zone" will also be wider and therefore larger mismatch is allowed.

Although the margin approach is originally applied to matched training-testing problems, it should also be effective in dealing with deviation of log likelihood values caused by noise distortion. In this paper, we apply the margin approach to improve the generalization capability of acoustic model for better robustness. We will describe how to maximize the margin for speech recognition in the next section.

B. Improving Generalization Capability by Maximizing the Margin

A large margin is the key to improve model's generalization capability. In [27, 28], SME was proposed to maximize the margin. In our experiments, we use SME to maximize the margin due to its good approximation of the expected risk. A brief description of SME is presented in this section. For detailed implementation and discussions about SME, please refer to [27].

In SME, the parameters of the acoustic model are estimated by minimizing an approximated expected risk as follows:

$$L^{SME}(\rho, \Lambda) = \frac{\lambda}{\rho} + R_{emp}(\rho, \Lambda) \tag{7}$$

where Λ is the set of acoustic model parameters, ρ is the soft margin, and $\frac{\lambda}{\rho}$ addresses the generalization risk. The variable λ is used to control the relative weights of the two items in (7). With a large λ, the training process will focus on reducing the generalization term and the margin will be large, and vice versa. To obtain good performance, it is important to obtain a good balance of these two terms.

The empirical risk is defined as the averaged risk of training utterances:

$$R_{emp}(\rho, \Lambda) = \frac{1}{N} \sum_{i=1}^{N} l(O_i, \rho, \Lambda) \tag{8}$$

where $O_i, i = 1, ..., N$ are the training utterances. The contribution of the utterance O_i to the total empirical risk is defined as:

$$l(O_i, \rho, \Lambda) = \begin{cases} \rho - d(O_i, \Lambda), & \text{if } \rho > d(O_i, \Lambda); \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $d(O_i, \Lambda)$ is a separation measure of O_i on model Λ. The separation measure usually represents how well the correct model is separated from competing models regarding O_i, or how far O_i is from the decision boundary. If the separation measure is not large enough, i.e. it is less than the margin, a loss is generated that equals to $\rho - d(O_i, \Lambda)$. In SME, the frame-normalized log likelihood ratio (LLR) defined in (5) is used as the separation measure.

The minimization of the objective function is solved by using generalized probabilistic descent (GPD) iteratively [28]. In order to obtain a differentiable loss function, the utterance loss function in (9) is embedded into a sigmoidal function as follows:

$$l(O_i, \rho, \Lambda) = \frac{\rho - d(O_i, \Lambda)}{1 + \exp(-\gamma(\rho - d(O_i, \Lambda)))} \tag{10}$$

where γ is used to control the transition slope of the sigmoidal function. With the smoothed loss function, the parameters of the acoustic model and the margin ρ can be jointly optimized iteratively:

$$\begin{cases} \Lambda_{t+1} = \Lambda_t - \eta_t \nabla L^{SME}(\rho, \Lambda)|_{\Lambda=\Lambda_t} \\ \rho_{t+1} = \rho_t - \kappa_t \nabla L^{SME}(\rho, \Lambda)|_{\rho=\rho_t} \end{cases} \tag{11}$$

where η_t and κ_t are the learning step size for acoustic model parameters and margin.

IV. EXPERIMENTS

A. System Description

In this section, we study the effect of improving model generalization capability on speech recognition performance for both matched and mismatched testing cases. The performance of SME is evaluated on Aurora-2 [29] and Aurora-3 [32] tasks. The acoustic models use standard "simple back-end" configurations, in which each digit is modeled by 16-state HMM with 3 Gaussian mixtures per state. MFCC features are used for system training and testing 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 [28]).

In our experimental study, we will compare SME with another popular discriminative training (DT) criterion, i.e. the minimum classification error (MCE) criterion [33–35]. Our purpose is to demonstrate the good characteristics of SME in improving model generalization capability rather than to carry out a comprehensive comparison of the two criteria. Hence, we will only show the comparison on selected test scenarios on Aurora-2 task. Similar to SME, the implementation of MCE is also based on GPD and N-best competing alignments (N=5) are used in MCE training. For comparison, only the closest competing alignment is used in SME, hence MCE actually uses more confusion information than SME. The parameters used in MCE are as follows: the sigmoid parameters $\theta = 0$ and $\gamma = 0.066$; the likelihood modification parameter $\eta = 0.066$ (in equation (13) of [34]). For brevity of the paper, we won't compare SME with other popular DT criterion such as maximum mutual information estimation (MMIE) that has been shown to deliver limited performance improvement in noisy speech recognition tasks [36].

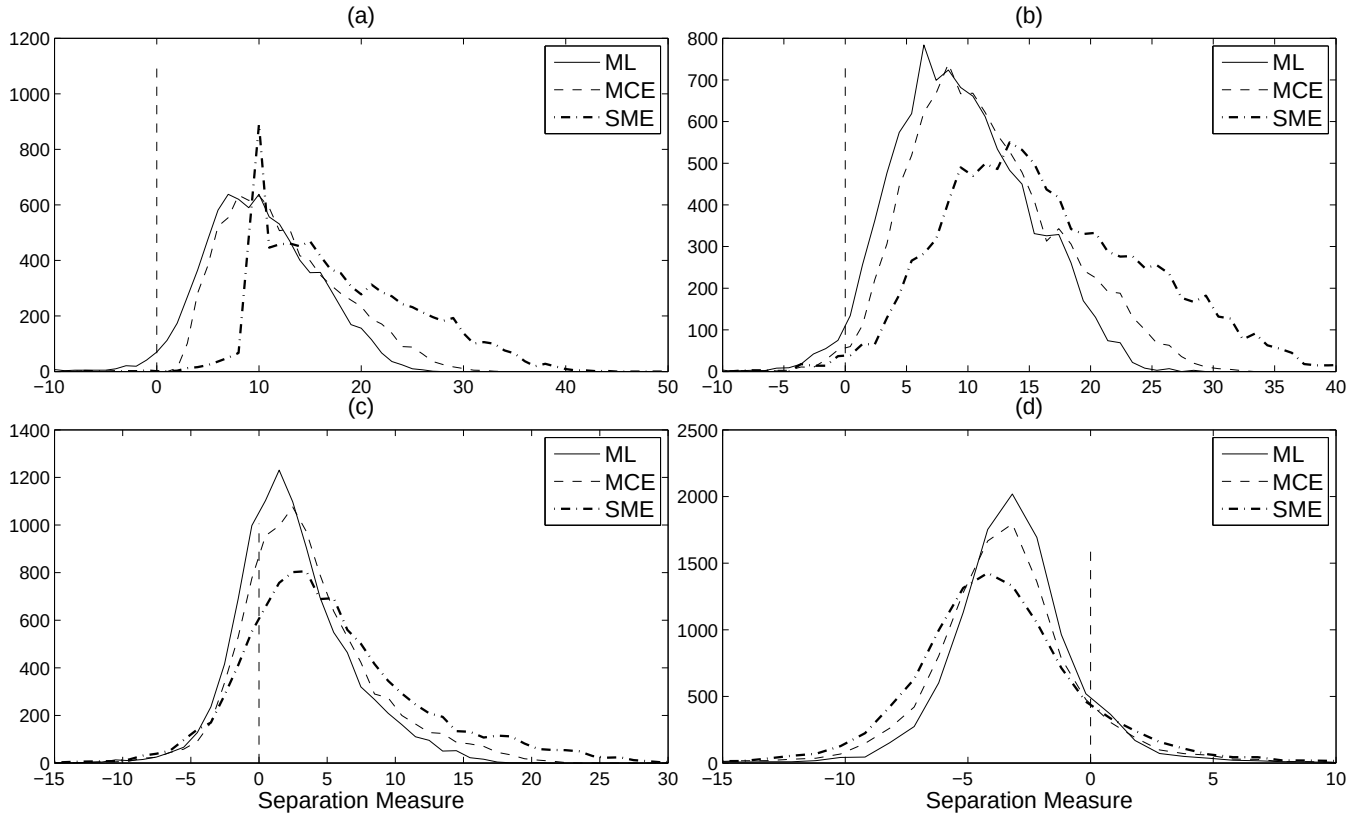

Fig. 4. Histograms of separation measures obtained by using ML and SME models on: (a) clean training data; (b) clean testing data; (c) 10dB test data; (d) -5dB test data.

B. Effect of SME on separation measures

Let's first examine how well SME improves the separation measure of the training and testing data. Note that a larger separation measure corresponds to a larger distance between a sample and the decision boundary and better separation. In Fig. 4, we compare the histograms of separation measures obtained using acoustic models trained by ML and SME. The acoustic models are trained using clean training data. The features are processed by MVN [10] in an utterance-by-utterance fashion.

In Fig. 4(a), the histograms of separation measures of clean training data are shown. There are 8440 training utterances in the training set, hence there are 8440 separation measures also. From the figure, we can see that the histogram obtained using SME is shifted right significantly. This indicates that the separation measures of training utterances are significantly improved compared to the ML baseline. Furthermore, there is a sharp slope around 9 in the histogram of SME. This is because when the training process stops, the final margin value is 9.14. By comparing the histograms of SME and ML, SME increases the separation measures of most training utterances to be larger than the margin. This observation indicates that in the log likelihood domain, the distance between most training utterances and their decision boundaries against the closest competing classes are larger or equal to the margin. After SME training, the portion of the histogram on the left side of the zero line is very small, which indicates a very small empirical risk, or training error. Compared to SME, the improvement of separation measures by MCE is quite limited. From the curves, SME allows the test data to deviate from the clean training data with longer distance than MCE in the log likelihood domain.

In Fig. 4(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. In the figure, we also observe significant increase of separation measures achieved by SME. However, we don't observe a sharp increase of separation measure as we do in Fig. 4(a). The SME also significantly reduces the amount of utterances whose separation measures are smaller than zero, i.e. wrongly classified utterances. Again, the improvement of MCE is less significant than that of SME.

In Fig. 4(c), the separation measure histograms of 10dB test sets are shown. From the figure, the effect of SME becomes less significant in 10dB test set than in clean test set. One reason is that the confusion pattern of noisy testing data may be different from that of clean training data. Therefore, what SME learns from clean training data becomes less relevant when the model is tested on lower SNR data. Note that the separation measures of most utterances evaluated on ML model (ML separation measures) are larger than -9. However, from the figure, only small portion of them are covered by SME trained acoustic model, in which there is a buffer zone with width=9.14. The majority of utterances that are wrongly classified by ML model are still wrongly classified by SME.

TABLE I
COMPARISON OF SME EFFECTS ON CORRECTLY AND WRONGLY CLASSIFIED UTTERANCES. CORRECT REFERS TO THOSE UTTERANCES CORRECTLY CLASSIFIED BY ML MODEL, AND WRONG REFERS TO THE REST UTTERANCES.

Group	Clean	20dB	15dB	10dB	5dB	0dB	-5dB
Correct	6.81	4.72	3.97	3.27	2.62	2.05	0.56
Wrong	3.81	2.23	1.68	1.29	0.68	-0.08	-0.71

This demonstrates the complexity of noise effect in speech recognition, which cannot be analyzed in a simple way.

In Fig. 4(d), the separation measure histograms of -5dB test sets are shown. In this SNR level, as the noise is more dominant than speech, SME actually decreases the mean of the histogram. However, the right tail part of the histogram is improved, and the number of correctly classified utterances is increased. The reason may be that SME is able to improve those relatively good utterances, while it degrades separation measures for those bad utterances. We will show next that SME performs differently for good and bad utterances. As compared to SME, MCE produces better separation measure than SME on the left of the vertical line x=0, and worse separation measure than SME on the right of x=0.

We also compare the SME effect on two groups of utterances, i.e. the group that is correctly classified by ML model and the group that is not. The comparison is shown in Table I, where the average absolute increases of separation measures achieved by SME over ML are shown. From the table, it is obvious that SME performs better for those utterances already correctly classified utterances by ML model, i.e. those relatively good utterances. The reason for this is similar to the reason for the different effects of SME at different SNR levels. For utterances in relatively better conditions, the deviation in log likelihood domain is smaller, the SME training is more relevant, hence, the effect of SME is more obvious.

C. Effect of Margin Size

An important question in SME training is the determination of the margin size. From our previous discussions, we expect that wide margin will make the acoustic model more general and robust. In this section, we will study the effect of SME on model training and speech recognition with different margin sizes.

In SME, the margin is not fixed, but jointly estimated with the acoustic model parameters by using the GPD algorithm. The variable λ is used to control the relative weights of the generalization term and the empirical risk term in the objective function of SME [see (7)]. Usually, larger λ will produce larger margin and more general model. We now study four λ values: 0.2, 1, 5, and 25. The acoustic model is trained from clean data and the features are processed by MVN.

The average recognition accuracies obtained by SME with different λ values are shown in Table II. From the table, it is observed that λ=0.2 produces poor performance, while the other three λ values produce similar results. For λ=0.2, SME improves recognition performance significantly at high SNR levels (clean, 20dB), but decreases performance at low SNR levels (5dB, 0dB, -5dB). For the other three λ values,

TABLE II
PERFORMANCE OF SME WITH MVN PROCESSED MFCC FEATURES WITH DIFFERENT λ VALUES ON AURORA-2 TASK. THE MODEL IS TRAINED FROM CLEAN DATA. ML REPRESENTS THE MAXIMUM LIKELIHOOD BASELINE.

SNR	ML	MCE	SME with different λ			
			0.2	1	5	25
Clean	99.16	99.58	99.43	99.64	**99.68**	99.64
20dB	97.42	98.40	97.83	98.50	**98.51**	98.41
15dB	95.17	96.66	95.44	**96.99**	96.85	96.66
10dB	89.34	91.98	89.60	93.05	**93.09**	92.70
5dB	74.48	79.26	74.28	82.85	**82.93**	82.43
0dB	45.21	51.60	43.41	58.33	58.67	**58.71**
-5dB	17.81	20.40	17.25	**26.55**	24.90	25.07
0-20dB	80.33	83.58	80.11	85.94	**86.01**	85.78
Margin	-	-	1.00	7.62	10.31	12.39

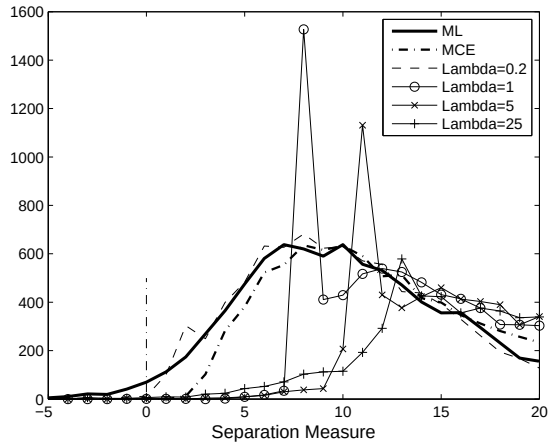

Fig. 5. Histogram of separation measures of training data with different λ values.

SME improves performance at all SNR levels. The last row of the table shows the margins estimated by SME when the accuracies shown in the table are obtained. As we expected, larger λ produces larger margin. Recognition results produced by MCE training are also shown for comparison. The performance improvement of MCE is less significant than the best performance improvement of SME at all SNR levels.

To examine the reason of the different performance shown in Table II, let's investigate the separation measures of training data. In Fig. 5, the histograms of separation measures of training data are shown with different λ values. These separation measures are obtained when the accuracies in Table II are obtained. Compared to ML, λ=0.2 does not change the separation measures very much, except that the number of training utterances whose separation measures are less than the margin (1.00 in this case) is reduced. Hence, the generalization capability of acoustic model with λ=0.2 is quite poor and this leads to poor recognition performance at mismatched testing scenarios when SNR level is low as shown in Table II. When using λ values of 1, 5, and 25, SME improves the separation measures of training data significantly, and larger λ produces bigger improvement. However, the difference in recognition performance of the three cases are quite insignificant. It is also observed that using λ values of 1, 5, and 25 all produce better separation measure histogram than MCE. This explains

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

SNR	Clean Condition			Multi-Condition		
	ML	SME	*Imp.*	ML	SME	*Imp.*
Clean	99.04	99.57	*55.06*	98.60	99.13	*37.89*
20dB	94.36	97.56	*56.69*	97.66	98.67	*43.11*
15dB	85.58	92.99	*51.35*	96.69	98.05	*41.17*
10dB	66.82	77.36	*31.77*	94.38	96.38	*35.66*
5dB	39.20	48.50	*15.30*	86.77	90.18	*25.81*
0dB	17.14	23.44	*7.60*	59.46	66.51	*17.39*
-5dB	9.78	11.70	*2.13*	24.27	26.65	*3.14*
0-20dB	60.62	67.97	***18.66***	86.99	89.96	***22.82***

the better performance of SME than MCE when λ is properly chosen in Table II.

D. Speech Recognition Performance on Aurora-2

We first examine the performance of SME with raw MFCC features. As the performance of SME is not very sensitive to the value of λ, we set λ to be 5 for all following speech recognition experiments. The MFCC features used here are not processed by any feature compensation methods.

The performance of SME on Aurora-2 task is shown in Table III. From the table, SME improves recognition accuracy significantly for both clean and multi-condition training schemes. This shows the effectiveness of our approach that improves model robustness by improving its generalization capability. There are some differences between the two training schemes in terms of relative error rate reduction. In clean condition training, SME performs better at high signal-to-noise ratio (SNR) levels (15dB and above) than at low SNR levels (5dB and below). This may be due to that the features at low SNR levels are too different from the clean training features, therefore, even more general model is not able to perform well. In multi-condition training, as the training data include noisy data down to 5dB, we see more even improvements at all SNR levels.

E. Interaction with MVN

As we have discussed in previous sections, one theoretical difficulty in applying SME to noisy speech recognition is the mismatch between training and testing distributions. In this section, we will study the effect of reducing training-testing mismatch on SME training. A simple and effective feature normalization method, MVN [10], is used to process both the training and testing features before model training and testing. Each dimension of the 39 MFCC features are processed by utterance-based MVN individually.

The performance of the combined system is shown in Table IV. From the table, we observe about 28% relative error rate reduction for both clean and multi-condition training. If we compare the results in Table IV and those in Table III, we can see that SME gains further performance when working with MVN in terms of average relative error rate reduction. For example, for clean condition training, as MVN reduces the mismatch between noisy test features and clean training features, SME produces higher improvement in low SNR levels (5dB and below). However, the relative error rate reduction in 20dB and 15dB are decreased. For multi-condition training, we see better performance of SME in all SNR levels.

TABLE IV
PERFORMANCE OF SME WITH MVN-PROCESSED FEATURES ON AURORA-2 TASK.

SNR	Clean Condition			Multi-Condition		
	ML	SME	*Imp.*	ML	SME	*Imp.*
Clean	99.16	99.68	*61.86*	98.23	99.20	*54.80*
20dB	97.42	98.51	*42.19*	98.53	99.28	*51.19*
15dB	95.17	96.85	*34.76*	97.70	98.93	*53.71*
10dB	89.34	93.09	*35.16*	96.09	97.92	*46.67*
5dB	74.48	82.93	*33.12*	90.71	94.02	*35.63*
0dB	45.21	58.67	*24.57*	74.26	79.28	*19.49*
-5dB	17.81	24.90	*8.63*	40.87	45.43	*7.70*
0-20dB	80.33	86.01	***28.89***	91.46	93.89	***28.42***

The experimental results show good interaction between SME and MVN. After MVN, the global mean and variance of both training and testing data become zero and one, respectively. Hence, the assumption of same distribution for both training and testing data in the statistical learning theory is less violated. We expect SME to work well with other feature domain methods as well.

F. Speech Recognition Performance on Aurora-3

We also evaluate our approach on Aurora-3 task, in which the data were recorded in real noisy environments. Our evaluations are based on raw MFCC and MVN-processed MFCC features. The value of λ is 5 and not tuned.

The performance of SME with raw MFCC features on Aurora-3 is shown in Table V. From the results, we have several observations. First, SME improves recognition accuracy for all cases except for the high-mismatch (HM) of German. This shows that by making the model more general, better performance can be obtained in realistic tasks. Second, SME usually produces higher relative error rate reduction in more matched cases. In most cases, the improvement for well-match (WM) is always the highest, followed by medium-mismatch (MM), and improvement is usually the lowest for HM. This is because in more mismatch training-testing cases, what SME learns from the training data is less relevant to the recognition of test data. The mismatch in HM may be beyond the generalization capability of the SME-trained acoustic model to tolerate. Similar results are observed in Aurora-2, 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. Similar to results in Table V, SME improves recognition accuracies significantly. This further manifested synergistic interaction between MVN and SME.

V. CONCLUSIONS

In this paper, we studied the effect of acoustic model generalization capability on robust speech recognition tasks.

TABLE V
PERFORMANCE OF SME WITH RAW MFCC FEATURES 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	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*
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	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*	ML	SME	*Imp.*
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***

Specifically, SME is used to increase the separation measures of training data to be larger than the margin, and therefore improve the generalization capability of the model. Experimental results confirmed that by making the acoustic model more general, speech recognition can tolerate certain level of mismatch between training and testing data. This shows another way of improving system robustness other than feature compensation and model adaptation methods. In addition, we observed that SME is more effective in modest mismatched scenarios than severe mismatched scenarios. Furthermore, we showed that feature domain method MVN works well together with SME, since MVN reduces the mismatch between the training and testing data such that SME-trained acoustic model is able to better tolerate the mismatch. We also expect better performance could be obtained when SME is combined with other noise robust methods. Currently, we are applying SME to large vocabulary tasks such as Aurora-4.

REFERENCES

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

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, Apr. 1985.

[4] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 275–296, 2004.

[5] M. Afify, "Accurate compensation in the log-spectral domain for noisy speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 388–398, 2005.

[6] L. Deng, J. Droppo, and A. Acero, "Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 133–143, Mar. 2004.

[7] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 218–223, May 2004.

[8] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large-vocabulary speech recognition under adverse acoustic environment," in *Proc. ICSLP '00*, (Beijing, China), pp. 806–809, Oct. 2000.

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

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

[11] 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.

[12] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[13] Y. Suh, M. Ji, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," *IEEE Signal Processing letters*, vol. 14, no. 4, pp. 287–290, 2007.

[14] J. C. Segura, C. Benítez, A. de la Torre, A. J. Rubio, and J. Ramírez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Processing letters*, vol. 11, no. 5, pp. 517–520, 2004.

[15] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[16] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.

[17] X. Xiao, E. S. Chng, and H. Li, "Temporal structure normalization of speech feature for robust speech recognition," *IEEE Signal Processing letters*, vol. 14, no. 7, pp. 500–503, 2007.

[18] X. Xiao, E. S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *to appear in the IEEE Trans. on Audio, Speech, and Language Processing*.

[19] 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, pp. 171–185, Apr. 1995.

[20] 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, pp. 291–298, Apr. 1994.

[21] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition," *Speech Communication*, vol. 12, pp. 231–239, Jul. 1993.

[22] Y. Tsao and C.-H. Lee, "An ensemble modeling approach to joint characterization of speaker and speaking environments," in *Proc. Eurospeech '07*, (Antwerp, Belgium), pp. 1050–1053, Sept. 2007.

[23] Y. Tsao and C.-H. Lee, "Two extensions to ensemble speaker and speaking environment modeling for robust automatic speech recognition," in *Proc. ASRU '07*, (Kyoto, Japan), pp. 77–80, Dec. 2007.

[24] 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.

[25] J. Li, L. Deng, D. Yu, Y. Gong, and A. Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive

distortions via vector taylor series," in *Proc. ASRU '07*, (Kyoto, Japan), pp. 65–70, Dec. 2007.
[26] V. N. Vapnik, *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
[27] 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.
[28] J. Li and C.-H. Lee, "On a generalization of margin-based discriminative training to robust speech recognition," in *Proc. InterSpeech '08*, (Brisbane, Australia), Sept. 2008.
[29] 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), pp. 29–32, Oct. 2000.
[30] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. PhD thesis, ECE, Carnegie Mellon University, 1990.
[31] P. J. Moreno, *Speech recognition in noisy environments*. PhD thesis, ECE, Carnegie Mellon University, 1996.
[32] Aurora document no. AU/255/00, *Baseline results for subsect of SpeechDat-Car Finnish database for ETSI STQ WI008 advance front end evaluation*, Nokia, Jan 2000.
[33] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Signal Processing*, vol. 40, pp. 3043–3054, Dec 1992.
[34] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.
[35] 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.
[36] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," in *Proc. InterSpeech '05*, (Lisbon, Portugal), pp. 989–992, Sept. 2005.