Noisy Speech Recognition Performance of Discriminative HMMs

Jun Du[1], Peng Liu[2], Frank Soong[2], Jian-Lai Zhou[2] and Ren-Hua Wang[1]

[1] University of Science and Technology of China, Hefei, 230027
[2] Microsoft Research Asia, Beijing, 100080
unuedjwj@ustc.edu, {pengliu,frankkps,jlzhou}@microsoft.com rhw@ustc.edu.cn

Abstract. Discriminatively trained HMMs are investigated in both clean and noisy environments in this study. First, a recognition error is defined at different levels including string, word, phone and acoustics. A high resolution error measure in terms of minimum divergence (MD) is specifically proposed and investigated along with other error measures. Using two speaker-independent continuous digit databases, Aurora2(English) and CNDigits (Mandarin Chinese), the recognition performance of recognizers, which are trained in terms of different error measures and using different training modes, is evaluated under different noise and SNR conditions. Experimental results show that discriminatively trained models performed better than the maximum likelihood baseline systems. Specifically, for MD trained systems, relative error reductions of 17.62% and 18.52% were obtained applying multi-training on Aurora2 and CNDigits, respectively.

Keywords : Noise Robustness, Minimum Divergence, Minimum Word Error, Discriminative Training

1 Introduction

With the progress of Automatic Speech Recognition (ASR), noise robustness of speech recognizers attract more and more attentions for practical recognition systems. Various noise robust technologies which can be grouped into three classes. 1. Feature domain approaches, which aim at noise resistant features, e.g., speech enhancement, feature compensation or transformation methods [1]; 2. Model domain approaches, e.g., Hidden Markov Model (HMM) decompensation [2], Parallel Model Combination (PMC) [3], which aim at modeling the distortion of features in noisy environments directly; 3. Hybrid approaches.

In the past decade, discriminative training has been shown quite effective in reducing word error rates of HMM based ASR systems in clean environment. In the first stage, sentence level discriminative criteria, including Maximum Mutual Information (MMI) [4, 5], Minimum Classification Error (MCE) [6], were proposed and proven effective. Recently, new criteria such as Minimum Word Error (MWE) and Minimum Phone Error (MPE) [7], which are based on fine error analysis at word or phone level, have achieved further improvement in recognition performance.

In [8–10], noise robustness investigation on sentence level discriminative criteria such as MCE, Corrective Training (CT) is reported. Hence, we are motivated to give a more complete investigation of noise robustness for genaral minimum error training.

From a unified viewpoint of error minimization, MCE, MWE and MPE are only different in error definition. String based MCE is based upon minimizing sentence error rate, while MWE is based on word error rate, which is more consistent with the popular metric used in evaluating ASR systems. Hence, the latter yields better word error rate, at least on the training set [7]. However, MPE performs slightly but universally better than MWE on testing set [7]. The success of MPE might be explained as follows: when refining acoustic models in discriminative training, it makes more sense to define errors in a more granular form of acoustic similarity. However, binary decision at phone label level is only a rough approximation of acoustic similarity.

Therefore, we propose to use acoustic dissimilarity to measure errors. Because acoustic behavior of speech units are characterized by HMMs, by measuring Kullback-Leibler Divergence (KLD) [11] between two given HMMs, we can have a physically more meaningful assessment of their acoustic similarity. Given sufficient training data, “ideal” ML models can be trained to represent the underlying distributions and then can be used for calculating KLDs.

Adopting KLD for defining errors, the corresponding training criterion is referred as Minimum Divergence (MD) [12]. The criterion possesses the following advantages: 1) It employs acoustic similarity for high-resolution error definition, which is directly related with acoustic model refinement; 2) Label comparison is no longer used, which alleviates the influence of chosen language model and phone set and the resultant hard binary decisions caused by label matching. Because of these advantages, MD is expected to be more flexible and robust.

In our work, MWE, which matches the evaluation metric, and MD, which focus on refining acoustic dissimilarity, are compared. Other issues related to robust discriminative training, including how to design the maximum likelihood baseline, and how to treat with silence model is also discussed.

Experiments were performed on Aurora2 [13], which is a widely adopted database for research on noise robustness, and CNDigits, a Chinese continuous digit database. We tested the effectiveness of discriminative training on different ML baseline and different noise environments.

The rest of paper is organized as follows. In section 2, issues on noise robustness of minimum error training will be discussed. In section 3, MD training will be introduced. Experimental results are shown and discussed in section 4. Finally in section 5, we give our conclusions.

2 Noise Robustness Analysis of Minimum Error Training

In this section, we will have a general discuss on the major issues we are facing in robust discriminative training.

2.1 Error Resolution of Minimum Error Training

In [7] and [12], various discriminative trainings in terms of their corresponding optimization measures are unified under the framework of minimum error training, where the objective function is an average of the recognition accuracies $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$ of all hypotheses weighted by the posterior probabilities. For conciseness, we consider single utterance case:

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \mid \boldsymbol{O}) \mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}}) \tag{1}$$

where $\boldsymbol{\theta}$ represents the set of the model parameters; $\boldsymbol{O}$ is a sequence of acoustic observation vectors; $\boldsymbol{W}_{\mathrm{r}}$ is the reference word sequence; $\boldsymbol{\mathcal{M}}$ is the hypotheses space; $P_{\boldsymbol{\theta}}(\boldsymbol{W} \mid \boldsymbol{O})$ is the generalized posterior probability of the hypothesis $\boldsymbol{W}$ given $\boldsymbol{O}$, which can be formulated as:

$$P_{\boldsymbol{\theta}}(\boldsymbol{W} \mid \boldsymbol{O}) = \frac{P_{\boldsymbol{\theta}}^{\kappa}(\boldsymbol{O} \mid \boldsymbol{W}) P(\boldsymbol{W})}{\sum_{\boldsymbol{W}' \in \boldsymbol{\mathcal{M}}} P_{\boldsymbol{\theta}}^{\kappa}(\boldsymbol{O} \mid \boldsymbol{W}') P(\boldsymbol{W}')} \tag{2}$$

where κ is the acoustic scaling factor.

The gain function $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$ is an *accuracy* measure of $\boldsymbol{W}$ given its reference $\boldsymbol{W}_{\mathrm{r}}$. In Table 1, comparison among several minimum error criteria are tabulated. In MWE training, $\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$ is word accuracy, which matches the commonly used evaluation metric of speech recognition. However, MPE has been shown to be more effective in reducing recognition errors because it provides a more precise measurement of word errors at the phone level. We can argue this point by advocating the final goal of discriminative training. In refining acoustic models to obtain better performance, it makes more sense to measure acoustic similarity between hypotheses instead of word accuracy. The symbol matching does not relate acoustic similarity with recognition. The measured errors can also be strongly affected by the phone set definition and language model selection. Therefore, acoustic similarity is proposed as a finer and more direct error definition in MD training.

Table 1. Comparison among criteria of minimum error training. ($\boldsymbol{P}_{\boldsymbol{W}}$: *Phone sequence corresponding to word sequence* $\boldsymbol{W}$; LEV(,): *Levenshtein distance between two symbol strings*; $|\cdot|$: *Number of symbols in a string.*)

Criterion	$\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$	Objective
String based MCE	$\delta(\boldsymbol{W} = \boldsymbol{W}_{\mathrm{r}})$	Sentence accuracy
MWE	$\lvert \boldsymbol{W}_{\mathrm{r}} \rvert - \mathrm{LEV}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}})$	Word accuracy
MPE	$\lvert P_{\boldsymbol{W}_{\mathrm{r}}} \rvert - \mathrm{LEV}(P_{\boldsymbol{W}}, P_{\boldsymbol{W}_{\mathrm{r}}})$	Phone accuracy
MD	$-D(\boldsymbol{W}_{\mathrm{r}} \parallel \boldsymbol{W})$	Acoustic similarity

Here we aim to seeking how criteria with different error resolution performs in the noisy environments. In our experiments, whole-word model, which is commonly used in digit tasks, is adopted. For the noisy robustness analysis, MWE which matches with the model type and evaluation metric of speech recognition, will compared with MD, which possesses the highest error resolution as shown in Table 1.

2.2 Different Training Modes

In noisy environments, various ML trained baseline can be designed. So the effectiveness of minimum error training with different training modes will be explored. In [13], two different sets of training, clean-training and multi-training, are used. In clean-training mode, only clean speech is used for training. Hence, when testing in noisy environments, there will be a mismatch. To alleviate this mismatch, multi-training, in which training set is composed of noisy speech with different SNRs, can be applied. But multi-training can only achieve a “global SNR” match. To achieve a “local SNR” match, we propose a SNR-based training mode. In our SNR-based training, each HMM set is trained using the speech with a specific SNR. A big HMM set is composed of all the SNR-based HMM sets. So there will be several SNR-based models for each digit. When testing, we will adopt the multi-pronunciation dictionary to output the digital label. SNR-based training can be considered as a high resolution acoustic modeling of multi-training. Illustration of three training modes is shown in Fig. 1.

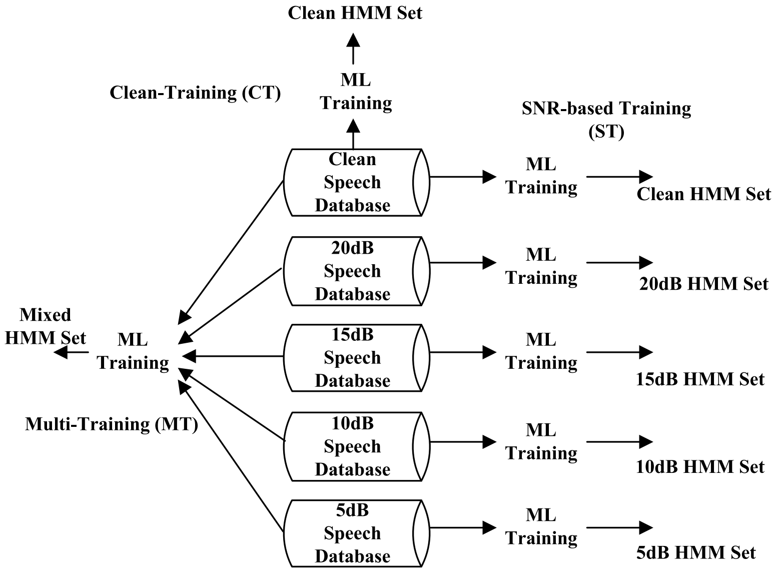

Fig. 1. Illustration of three training modes

2.3 Silence Model Update

Silence or background model can have a significant effect on word errors. Hence, whether or not to update silence model in minimum error training can be critical under noisy conditions. In our research, we pay special attention to this issue for reasonable guidelines.

3 Word Graph based Minimum Divergence Training

3.1 Defining Errors by Acoustic Similarity

A word sequence is acoustically characterized by a sequence of HMMs. For automatically measuring acoustic similarity between $\boldsymbol{W}$ and $\boldsymbol{W}_{\mathrm{r}}$, we adopt KLD between the corresponding HMMs:

$$\mathcal{A}(\boldsymbol{W}, \boldsymbol{W}_{\mathrm{r}}) = -D(\boldsymbol{W}_{\mathrm{r}} \parallel \boldsymbol{W}) \tag{3}$$

The HMMs, when they are reasonably well trained in ML sense, can serve as succinct descriptions of data.

3.2 KLD between Two Word Sequences

Given two word sequences $\boldsymbol{W}_{\mathrm{r}}$ and $\boldsymbol{W}$ without their state segmentations, we should use a state matching algorithm to measure the KLD between the corresponding HMMs [14]. With state segmentations, the calculation can be further decomposed down to the state level:

$$\begin{aligned} D(\boldsymbol{W}_{\mathrm{r}} \parallel \boldsymbol{W}) &= D(\boldsymbol{s}_{\mathrm{r}}^{1:T} \parallel \boldsymbol{s}^{1:T}) \\ &= \int p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}_{\mathrm{r}}^{1:T}) \log \frac{p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}_{\mathrm{r}}^{1:T})}{p(\boldsymbol{o}^{1:T} \mid \boldsymbol{s}^{1:T})} d\boldsymbol{o}^{1:T} \end{aligned} \tag{4}$$

where T is the number of frames; $\boldsymbol{o}^{1:T}$ and $\boldsymbol{s}_{\mathrm{r}}^{1:T}$ are the observation sequence and hidden state sequence, respectively.

By assuring all observations are independent, we obtain:

$$D(\boldsymbol{s}_{\mathrm{r}}^{1:T} \parallel \boldsymbol{s}^{1:T}) = \sum_{t=1}^{T} D(s_{\mathrm{r}}^{t} \parallel s^{t}) \tag{5}$$

which means we can calculate KLD state by state, and sum them up.

Conventionally, each state s is characterized by a Gaussian Mixture Model (GMM): $p(\boldsymbol{o} \mid s) = \sum_{m=1}^{M_s} \omega_{sm} \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{sm}, \boldsymbol{\Sigma}_{sm})$, so the comparison is reduced to measuring KLD between two GMMs. Since there is no closed-form solution, we need to resort to the computationally intensive Monte-Carlo simulations. The unscented transform mechanism [15] has been proposed to approximate the KLD measurement of two GMMs.

Let $\mathcal{N}(\boldsymbol{o};\boldsymbol{\mu},\boldsymbol{\Sigma})$ be a N-dimensional Gaussian distribution and h is an arbitrary $\mathbb{R}^N \rightarrow \mathbb{R}$ function, unscented transform mechanism suggests approximating the expectation of h by:

$$\int \mathcal{N}(\boldsymbol{o};\boldsymbol{\mu},\boldsymbol{\Sigma})h(\boldsymbol{o})d\boldsymbol{o} \approx \frac{1}{2N}\sum_{k=1}^{2N} h(\boldsymbol{o}_k) \tag{6}$$

where $\boldsymbol{o}_k(1 \le k \le 2N)$ are the artificially chosen "*sigma*" points: $\boldsymbol{o}_k = \boldsymbol{\mu} + \sqrt{N\lambda_k}\boldsymbol{u}_k$, $\boldsymbol{o}_{k+N} = \boldsymbol{\mu} - \sqrt{N\lambda_k}\boldsymbol{u}_k (1 \le k \le N)$, where λ_k, $\boldsymbol{u}_k$ are the k^{th} eigenvalue and eigenvector of $\boldsymbol{\Sigma}$, respectively. Geometrically, all these "*sigma*" points are on the principal axes of $\boldsymbol{\Sigma}$. Eq. 6 is precise if h is quadratic.

Based on Eq. 6, KLD between two Gaussian mixtures is approximated by:

$$D(s_\mathrm{r} \,\|\, s) \approx \frac{1}{2N}\sum_{m=1}^{M}\omega_m \sum_{k=1}^{2N} \log\frac{p(\boldsymbol{o}_{m,k} \,|\, s_\mathrm{r})}{p(\boldsymbol{o}_{m,k} \,|\, s)} \tag{7}$$

where $\boldsymbol{o}_{m,k}$ is the k^{th} "*sigma*" point in the m^{th} Gaussian kernel of $p(\boldsymbol{o}_{m,k} \,|\, s_\mathrm{r})$. By plugging it into Eq. 4, we obtain the KLD between two word sequences given their state segmentations.

3.3 Gain Function Calculation

Usually, word graph is a compact representation of large hypotheses space in speech recognition. Because the KLD between a hypothesised word sequence and the reference can be decomposed down to the frame level, we have the following word graph based representation of (1):

$$\mathcal{F}(\boldsymbol{\theta}) = \sum_{w \in \boldsymbol{\mathcal{M}}} \sum_{\boldsymbol{W} \in \boldsymbol{\mathcal{M}} : w \in \boldsymbol{W}} P_{\boldsymbol{\theta}}(\boldsymbol{W} \,|\, \boldsymbol{O})\mathcal{A}(w) \tag{8}$$

where $\mathcal{A}(w)$ is the gain function of word arc w. Denote b_w, e_w the start frame index and end frame index of w, we have:

$$\mathcal{A}(w) = -\sum_{t=b_w}^{e_w} D(s_w^t \,\|\, s_\mathrm{r}^t) \tag{9}$$

where the s_w^t and s_r^t represent the certain state at time t on arc w and the reference, respectively.

As mentioned in [7], we use Forward-Backward algorithm to update the word graph and the Extended Baum-Welch algorithm to update the model parameters in the training iterations.

4 Experiments

4.1 Experimental Setup

Experiments on both English (TIDigits and Aurora2) and Chinese (CNDigits) continuous digit tasks were performed. The English vocabulary is made of the

11 digits, from 'one(1)' to 'nine(9)', plus 'oh(0)' and 'zero(0)'. The Chinese vocabulary is made of digits from 'ling(0)' to 'jiu(9)', plus 'yao(1)'. The baseline configuration for three systems is listed in Table 2.

For TIDigits Experiments, man, woman, boy and girl speakers, were used in both training and testing.

The Aurora2 task consists of English digits in the presence of additive noise and linear convolutional channel distortion. These distortions have been synthetically introduced to clean TIDigits data. Three testing sets measure performance against noise types similar to those seen in the training data (set A), different from those seen in the training data (set B), and with an additional convolutional channel (set C). The baseline performance and other details can be found in [13].

The original clean database of CNDigits is collected by Microsoft Research Asia. 8 types of noises, i.e. waiting room of a station, platform, shop, street, bus, airport lounge, airport exit, outside, are used for noise addition. 8000 clean utterances from 120 female and 200 male speakers for training set are split into 20 subsets with 400 utterances in each subset. Each subset contains a few utterances of all training speakers. The 20 subsets represent 4 different noise scenarios at 5 different SNRs. The 4 noises are waiting room, street, bus and airport lounge. The SNRs are 20dB, 15dB, 10dB, 5dB and the clean condition. Two different test sets are defined. 3947 clean utterances from 56 female and 102 male speakers are split into 4 subsets with about 987 utterances in each. All speakers are involved in each subset. One noise is added to each subset at SNRs of 20dB, 15dB, 10dB, 5dB, 0dB, -5dB and the clean condition. In the first test set, called test set WM(Well-Match), the four noises, the same as those used in training set, are added to the 4 subset. The second test set, called test set MM(Mis-Match), is created in exactly the same way, but using four different noises , namely platform, shop, airport exit and outside. Our design of CNDigits database is similar to Aurora2.

For mininum error training, the acoustic scaling factor κ was set to $\frac{1}{33}$. All KLDs between any two states were precomputed to make the MD training more efficient. For Aurora2 and CNDigits, we select the best results after 20 iterations for each sub set of testing.

Table 2. Baseline configuration

System	Feature	Model Type	# State /Digit	# Gauss /State	# string of training set	# string of testing set
TIDigits	MFCC_E_D_A	left-to-right whole-word model without skipping	10	6	12549	12547
Aurora2			16	3	8440*2	1001*70
CNDigits			10	3	8000	987*56

4.2 Experiments on TIDigits Database

As a preliminary of noise robustness analysis, we first give the results of MD on the clean TIDigits database compared with MWE. As shown in Fig. 2, performance of MD achieves 57.8% relative error reduction compared with ML baseline and also outperforms MWE in all iterations.

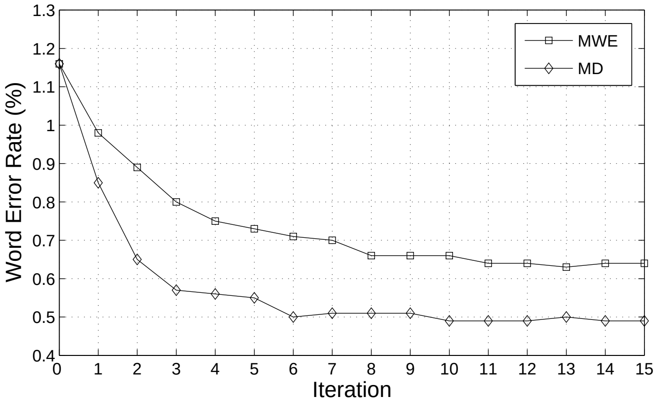

Fig. 2. Performance comparison on TIDigits

4.3 Experiments on Aurora2 Database

Table 3. Word Accuracy (%) of MWE with or without silence model update in different training modes on Aurora2.

Training Mode	Update Silence Model	Set A	Set B	Set C	Overall
Clean	YES	61.85	56.94	66.26	60.77
Clean	NO	64.74	61.69	67.95	64.16
Multi	YES	89.15	89.16	84.66	88.26
Multi	NO	88.91	88.55	84.43	87.87

Silence Model Update. As shown in Table 3, we explore whether to update silence model in minimum error training using different training modes. Because it is unrelated with criteria, here we adopt MWE. when applying clean-training, the performances on all test sets without updating silence model are consistently better. But in multi-training, the conclusion is opposite. From the results, we can conclude that increasing the discrimination of silence model will lead to performance degradation in mismatched cases (clean-training) and performance improvement in matched cases (multi-training). Obviously our SNR-based training belongs to the latter. In all our experiments, the treatment of silence model will obey this conclusion.

Table 4. Performance comparison on Aurora2 (MD vs. MWE)

Multi-Training - Results (Minimum Divergence)														
	A					B					C			Rel
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Impr
Clean	99.14	99.12	98.9	99.2	99.09	99.14	99.12	98.9	99.2	99.09	98.89	98.85	98.87	99.05	35.32%
20 dB	98.71	98.55	98.81	98.61	98.67	98.43	98.37	98.57	98.89	98.57	98.65	97.64	98.15	98.52	43.92%
15 dB	98.5	98	98.33	97.93	98.19	98	97.76	97.79	97.93	97.87	97.88	96.74	97.31	97.89	42.04%
10 dB	97.18	96.55	97.2	96.08	96.75	96.41	95.8	96.06	95.31	95.90	95.15	94.04	94.60	95.98	34.81%
5 dB	92.39	89.81	90.49	90.25	90.74	89.28	87.06	90.52	87.23	88.52	84.68	82.56	83.62	88.43	20.78%
0 dB	72.8	64.63	58.93	70.32	66.67	65.24	64	69.19	62.48	65.23	49.25	54.44	51.85	63.13	10.51%
-5dB	31.04	29.56	22.7	28.57	27.97	30.06	28.96	33.58	25.46	29.52	22.01	24.24	23.13	27.62	4.15%
Average	91.92	89.51	88.75	90.64	90.20	89.47	88.60	90.43	88.37	89.22	85.12	85.08	85.10	88.79	
Rel Impr	28.10%	12.93%	16.53%	21.79%	19.60%	27.93%	12.04%	22.53%	22.40%	21.45%	11.21%	4.93%	8.17%		17.62%

Multi-Training - Results (Minimum Word Error)														
	A					B					C			Rel
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average	Impr
Clean	99.14	99.18	99.02	99.29	99.16	99.14	99.18	99.02	99.29	99.16	98.99	99.06	99.03	99.13	40.96%
20 dB	98.86	98.67	98.78	98.7	98.75	98.74	98.43	98.72	98.95	98.71	98.34	97.4	97.87	98.56	45.45%
15 dB	98.74	98.13	98.33	97.69	98.22	98.5	97.82	98.03	98.06	98.10	97.33	96.25	96.79	97.89	41.97%
10 dB	96.87	95.95	96.87	95.43	96.28	96.22	95.53	96.42	95.74	95.98	94.63	93.5	94.07	95.72	30.03%
5 dB	92.32	88.85	88.25	88.83	89.56	88.36	87.3	89.53	86.61	87.95	84.49	82.62	83.56	87.72	15.40%
0 dB	70.31	63.33	53.44	64.7	62.95	64.6	68.18	68.27	59.12	65.04	47.62	54.44	51.03	61.40	6.25%
-5dB	29.66	29.72	21.8	25.27	26.61	30.21	27.84	33.49	23.97	28.88	21.31	24.24	22.78	26.75	3.01%
Average	91.42	88.99	87.13	89.07	89.15	89.28	89.45	90.19	87.70	89.16	84.48	84.84	84.66	88.26	
Rel Impr	23.69%	8.60%	4.53%	8.69%	10.98%	26.64%	18.62%	20.65%	17.92%	21.02%	7.39%	3.39%	5.46%		13.71%

Error Resolution of Minimum Error Training. As shown in Table 4, the performances of MD and MWE are compared. Here multi-training is adopted because it's believed that matching between training and testing can tap the potential of minimum error training. For the overall performance on three test sets, MD consistently outperforms MWE. From the viewpoint of SNRs, MD outperforms MWE in most cases when SNR is below 15dB. Hence, we can conclude that although MWE matches with the model type and evaluation metric of speech recognition, MD which possesses the highest error resolution outperforms it in low SNR. In other words, the performance can be improved in low SNR by increasing the error resolution of criterion in minimum error training.

Different Training Modes. Fig. 3 shows relative improvement over ML baseline using MD training with different training modes. From this figure, some

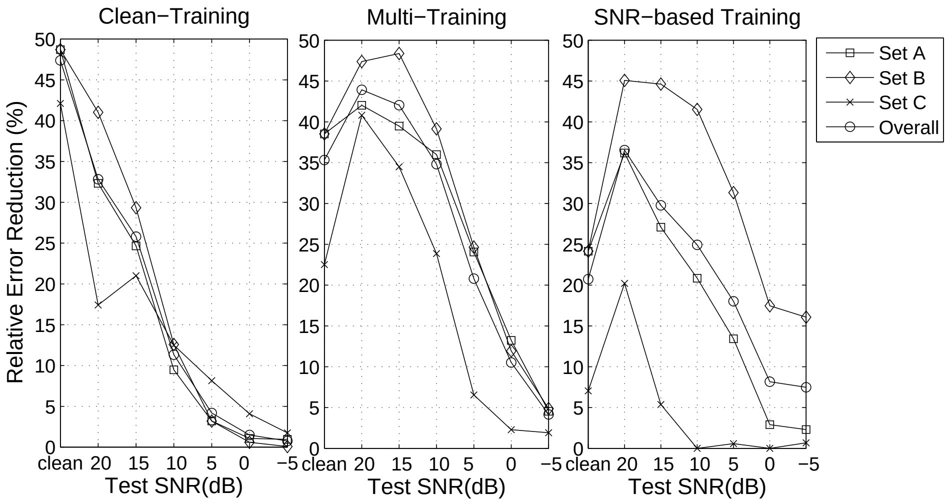

Fig. 3. Relative Improvement over ML baseline on Aurora2 using different training modes in MD training

Table 5. Summary of performance on Aurora2 using different training modes in MD training.

	Word Accuracy (%)				Relative Improvement			
Training Mode	Set A	Set B	Set C	Overall	Set A	Set B	Set C	Overall
Clean-Training	63.49	58.94	68.96	62.76	5.56%	7.21%	8.32%	6.76%
Multi-Training	90.20	89.22	85.10	88.79	19.60%	21.45%	8.17%	17.62%
SNR-based Training	91.27	89.27	86.70	89.56	10.00%	26.21%	1.14%	15.68%

conclusions can be obtained. First, Set B, whose noise scenarios are different from training achieves the most obvious relative improvement in most cases. The relative improvement of set A are comparable with set B in the clean-training and multi-training but worse than set B in SNR-based training. The relative improvement of set C, due to the mismatch of noise scenario and channel, almost the worst in all training modes. Second, the relative improvement performance declines for decreasing SNR in clean-training. But in multi-training and SNR-based training, the peak performance is in the range of 20dB to 15dB. Also in the low SNRs, the performance of cleaning-training is worse than the other two training modes on set A and set B.

The summary of performance is listed in Table 5. Word accuracy of our SNR-based training outperforms multi-training on all test sets, especially set A and set C. For the overall relative improvement, the best result of 17.62% is achieved in multi-training.

Table 6. Performance comparison on Chinese digit database (CNDigits) using multi-training

Multi-Training - Results (ML Reference)											
	Well-Matched(WM)					Mis-Matched(MM)					
	Waiting Room	Street	Bus	Lounge	Average	Platform	Shop	Outside	Exit	Average	Average
Clean	95.9	95.46	95.57	95.33	95.57	95.9	95.46	95.57	95.33	95.57	95.57
20 dB	95.54	95.35	95.56	95.07	95.38	95.93	94.99	95.55	95.17	95.41	95.40
15 dB	94.21	95.29	95.61	94.63	94.94	95.45	93.83	95.33	94.88	94.87	94.90
10 dB	91.15	94.17	95.52	93.82	93.67	94.12	90.4	94.2	94.25	93.24	93.45
5 dB	82.33	92.21	95.42	89.64	89.90	89.64	80.97	90.63	91.34	88.15	89.02
0 dB	65.42	84.63	94.85	77.43	80.58	77.4	64.46	80.77	82.36	76.25	78.42
-5dB	39.23	68.18	93.34	51.1	62.96	51.3	39.53	57.78	58.7	51.83	57.40
Average	85.73	92.33	95.39	90.12	90.89	90.51	84.93	91.30	91.60	89.58	90.24

Multi-Training - Results (Minimum Word Error)												
	Well-Matched(WM)					Mis-Matched(MM)						Rel
	Waiting Room	Street	Bus	Lounge	Average	Platform	Shop	Outside	Exit	Average	Average	Impr
Clean	96.95	96.45	96.74	96.52	96.67	96.95	96.45	96.74	96.52	96.67	96.67	24.83%
20 dB	96.19	96.33	96.74	96.26	96.38	96.73	95.96	96.66	96.3	96.41	96.40	21.72%
15 dB	94.84	96.06	96.78	95.76	95.86	95.95	94.69	96.12	95.96	95.68	95.77	17.23%
10 dB	92.32	94.84	96.72	94.39	94.57	94.53	91.62	94.99	94.73	93.97	94.27	12.80%
5 dB	86.09	92.66	96.47	90.53	91.44	90.07	84.5	91.91	92.05	89.63	90.54	12.89%
0 dB	70.89	85.06	95.99	78.82	82.69	78.72	67.74	82.18	83.47	78.03	80.36	9.45%
-5dB	42.58	69.2	94.33	51.89	64.50	51.72	40.46	58.99	59.48	52.66	58.58	4.04%
Average	88.07	92.99	96.54	91.15	92.19	91.20	86.90	92.37	92.50	90.74	91.47	
Rel Impr	16.37%	8.60%	24.91%	10.46%	14.21%	7.29%	13.09%	12.36%	10.74%	11.14%		12.57%

Multi-Training - Results (Minimum Divergence)												
	Well-Matched(WM)					Mis-Matched(MM)						Rel
	Waiting Room	Street	Bus	Lounge	Average	Platform	Shop	Outside	Exit	Average	Average	Impr
Clean	97.21	96.67	97.06	96.59	96.88	97.21	96.67	97.06	96.59	96.88	96.88	29.80%
20 dB	96.19	96.46	96.98	96.29	96.48	96.92	96.1	96.84	96.38	96.56	96.52	24.46%
15 dB	95.02	96.35	97.08	95.97	96.11	96.26	94.77	96.44	96.11	95.90	96.00	21.97%
10 dB	92.44	95.02	96.87	94.64	94.74	94.78	91.93	95.29	94.92	94.23	94.49	16.27%
5 dB	86.47	92.75	96.77	90.87	91.72	90.99	85.48	92.4	92.5	90.34	91.03	17.59%
0 dB	72.32	85.95	96.17	81.51	83.99	81.78	69.78	84.63	85.4	80.40	82.19	17.99%
-5dB	46.63	72.31	94.5	57.75	67.80	58.93	43.53	64.83	64.48	57.94	62.87	13.64%
Average	88.49	93.31	96.77	91.86	92.61	92.15	87.61	93.12	93.06	91.49	92.05	
Rel Impr	19.33%	12.72%	29.99%	17.59%	18.81%	17.26%	17.80%	20.96%	17.40%	18.25%		18.52%

4.4 CNDigits Database Experiments

On CNDigits database, we compare MD and MWE with ML applying multi-training as a further verification of conclusions on Aurora2. Performances are shown in Table 6. Totally MD achieves 18.52% relative improvement over ML baseline. Although minimum error training on both English and Chinese is effective in noisy envrionments, there are still some differences. First, the most obvious relative improvement on CNDigits occurs in clean condition which is different from that on Aurora2. Second, more than 10% relative improvement is still obtained at low SNRs (below 0dB) on CNDigits. Third, MD outperforms MWE in all noisy conditions.

5 Conclusions

In this paper, the noise robustness of discriminative training is investigated. Discriminatively trained models are tested on both English and Chinese continuous digit databases in clean and noisy conditions. Most experiments adopt MD criterion. First, silence model should only be updated when the training and testing data are matched (Both are noisy data). Second, minimum error training is effective in noisy conditions for both clean-training and multi-training, even for SNR-based training which produces higher resolution acoustic models. Third, MD with higher error resolution than MWE is more robust in low SNR scenarios. Even when testing on mismatched noise scenarios, minimum error training is also noise robust as matched noise scenarios.

In future work, we will focus on seeking noise resistant features based on minimum error training and improve performance further in noise conditions.

References

1. Gong, Y.: Speech Recognition in Noisy Environments: A Survey. Speech Communication, Vol.16. (1995) 261-291
2. Varga, A.P., Moore, R.K.: Hidden Markov model decomposition of speech and noise. Proc. ICASSP (1990) 845-848
3. Gales, M.J.F., Young, S.J.: Robust Continuous Speech Recognition using Parallel Model Combination. Tech.Rep., Cambridge University (1994)
4. Schluter, R.: Investigations on Discriminative Training Criteria. Ph.D.thesis, Aachen University (2000)
5. Valtchev, V., Odell, J.J., Woodland, P.C., Young, S.J.: MMIE Training of Large Vocabulary Speech Recognition Systems. Speech Communication, Vol.22. 303-314
6. Juang, B.-H., Chou, W., Lee, C.-H.: Minimum Classification Error Rate Methods for Speech Recogtion. IEEE Trans. on Speech and Audio Processing, Vol.5. No.3.(1997) 257-265
7. Povey, D.: Discriminative Training for Large Vocabulary Speech Recognition. Ph.D. Thesis, Cambridge University (2004)
8. Ohkura, K., Rainton, D., Sugiyama, M.: Noise-robust HMMs Based on Minimum Error Classification. Proc.ICASSP (1993) 75-78
9. Meyer, C., Rose, G.: Improved Noise Robustness by Corrective and Rival Training. Proc.ICASSP (2001) 293-296
10. Laurila, K., Vasilache, M., Viikki, O.: A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition. Proc.ICASSP (1998) 85-88
11. Kullback, S., Leibler, R.A.: On Information and Sufficiency. Ann. Math. Stat, Vol. 22. (1951) 79-86
12. Du, J., Liu, P., Soong, F.K., Zhou, J.-L., Wang, R.H.: Minimum Divergence Based Discriminative Training. Accepted by Proc.ICSLP (2006)
13. Hirsch, H.G., Pearce, D.: The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions. In ISCA ITRW ASR2000, Paris France (2000)
14. Liu, P., Soong, F.K., Zhou, J.-L.: Effective Estimation of Kullback-Leibler Divergence between Speech Models. Tech.Rep., Microsoft Research Asia (2005)
15. Goldberger, J.: An Efficient Image Similarity Measure based on Approximations of KL-Divergence between Two Gaussian Mixtures. Proc. International Conference on Computer Vision 2003, Nice France (2003) 370-377