Computational Linguistics and Chinese Language Processing
Vol. 12, No. 3, September 2007, pp. 291-302

© The Association for Computational Linguistics and Chinese Language Processing

Performance of Discriminative HMM Training in Noise

Jun Du*,+, Peng Liu+, Frank K. Soong+, Jian-Lai Zhou+, and

Ren-Hua Wang*

Abstract

In this study, discriminative HMM training and its performance are investigated in both clean and noisy environments. Recognition error is defined at string, word, phone, and acoustic levels and treated in a unified framework in discriminative training. With an acoustic level, high-resolution error measurement, a discriminative criterion of minimum divergence (MD) is proposed. Using speaker-independent, continuous digit databases, Aurora2, the recognition performance of recognizers, which are trained in terms of different error measures and different training modes, is evaluated under various noise and SNR conditions. Experimental results show that discriminatively trained models perform better than the maximum likelihood baseline systems. Specifically, in MWE and MD training, relative error reductions of 13.71% and 17.62% are obtained with multi-training on Aurora2, respectively. Moreover, compared with ML training, MD training becomes more effective as the SNR increases.

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 attracts more and more attention for practical recognition systems. Various noise robust technologies 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 [Gong 1995]; 2. Model domain approaches, *e.g.*, Hidden Markov

* University of Science and Technology of China, Hefei, P. R. China, 230027
Tel: +86-551-3601363-806 Fax: +86-551-3601363-807
E-mail: unuedjwj@ustc.edu; rhw@ustc.edu.cn
+ Microsoft Research Asia, Beijing, P. R. China, 100080
E-mail: {pengliu, frankkps, jlzhou }@microsoft.com

[Received March 31, 2007; Revised August 14, 2007; Accepted August 17, 2007]

Model (HMM) decompensation [Varga *et al.* 1990], Parallel Model Combination (PMC) [Gales *et al.* 1994], 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 a clean environment. In the first stage, sentence level discriminative training criteria, including Maximum Mutual Information (MMI) [Schluter 2000; Valtchev *et al.* 1997] and Minimum Classification Error (MCE) [Juang *et al.* 1997], were proposed and proven effective. Recently, new criteria such as Minimum Word Error (MWE) and Minimum Phone Error (MPE) [Povey 2004], which are based on fine error analysis at word or phone level, have achieved further improvement in recognition performance.

In [Ohkura *et al.* 1993; Meyer *et al.* 2001; Laurila *et al.* 1998], noise robustness investigation on sentence level discriminative criteria such as MCE, Corrective Training (CT) is reported. Hence, we give a more complete investigation of noise robustness for general minimum error training.

From a unified view of error minimization, the major difference between MCE, MWE and MPE is the 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 a better word error rate, at least on the training set [Povey 2004]. However, MPE performs slightly but universally better than MWE on the testing set [Povey 2004]. 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, error definition at phone label level is only a rough approximation of acoustic similarity.

Based on the analysis above, we have proposed using acoustic dissimilarity to measure errors [Du *et al.* 2006]. As acoustic behavior of speech units is characterized by HMMs, by measuring Kullback-Leibler Divergence (KLD) [Kullback *et al.* 1951] between two given HMMs, we can obtain a physically more meaningful assessment of their acoustic similarity.

Adopting KLD for defining dissimilarity, the corresponding training criterion is referred as Minimum Divergence (MD) [Du *et al.* 2006; Du *et al.* 2007]. The criterion possesses the following potential advantages: 1) It employs acoustic similarity for high-resolution error definition, which is directly related to acoustic model refinement; 2) Label comparison is no longer used, which alleviates the influence of the chosen language model and phone set and the resultant hard binary decisions caused by label matching. Due to these advantages, MD is expected to be more flexible and robust.

In our work, MWE, which matches the evaluation metric, and MD, which focuses 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 the silence model is also discussed.

Experiments were performed on Aurora2 [Hirsch *et al.* 2000], which is a widely adopted database for research on noise robustness. For completeness, we tested the effectiveness of discriminative training on different ML baselines 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 give a brief discussion of the major issues we are facing in robust discriminative training.

2.1 Error Resolution of Minimum Error Training

In [Povey 2004] and [Du *et al.* 2006], various discriminative training approaches 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 the single training utterance case:

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

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; $\mathcal{M}$ is the hypotheses space; $P_{\theta}(\boldsymbol{W}\mid\boldsymbol{O})$ is the posterior probability of the hypothesis $\boldsymbol{W}$ given $\boldsymbol{O}$, which can be formulated as:

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

where κ is the acoustic scaling factor.

The gain function $\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_{r})$ is an *accuracy* measure of $\boldsymbol{W}$ given its reference $\boldsymbol{W}_{r}$. In Table 1, comparison of several minimum error criteria are listed. In MWE training, $\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_{r})$ 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 of criteria of minimum error training. (P_W : Phone sequence corresponding to word sequence W; LEV(,): Levenshtein distance between two symbol strings;$|\cdot|$: Number of symbols in a string.)

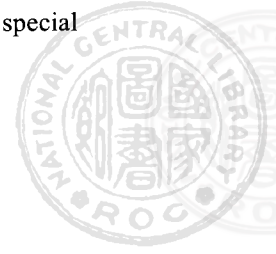
Criterion	$A(W, W_r)$	Objective
String based MCE	$\delta(W = W_r)$	Sentence accuracy
MWE	$\lvert W_r \rvert - \mathrm{LEV}(W, W_r)$	Word accuracy
MPE	$\lvert P_{W_r} \rvert - \mathrm{LEV}(P_W, P_{W_r})$	Phone accuracy
MD	$-D(W_r \parallel W)$	Acoustic similarity

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

2.2 Training Modes

In noisy environments, various ML trained baselines can be designed. So, the effectiveness of minimum error training with different training modes will be explored. In [Hirsch *et al.* 2000], two different sets of training, clean-training and multi-training, are used. In clean-training mode, only clean speech is used for training. Hence, there will be a mismatch when the model is tested in noisy environments. To alleviate the mismatch, multi-training, in which training set is composed of noisy speech with different SNRs, can be applied. Actually, multi-training can only achieve a "global SNR" match. To achieve a "local SNR" match, we adopt a SNR-based training mode. In the training phase, we train a series of models at different SNR levels, while in testing, all these models are paralleled as multi pronunciations of a HMM. Ideally, the model that matched the local SNR best will be automatically selected in decoding. SNR-based training can be considered as a high resolution acoustic modeling of multi-training. An illustration of the three training modes is shown in Figure 1.

An important issue in discriminative training is how to update silence or background models, which is even more critical in a noisy environment. In our research, we pay special attention to this issue for appropriate guidelines.



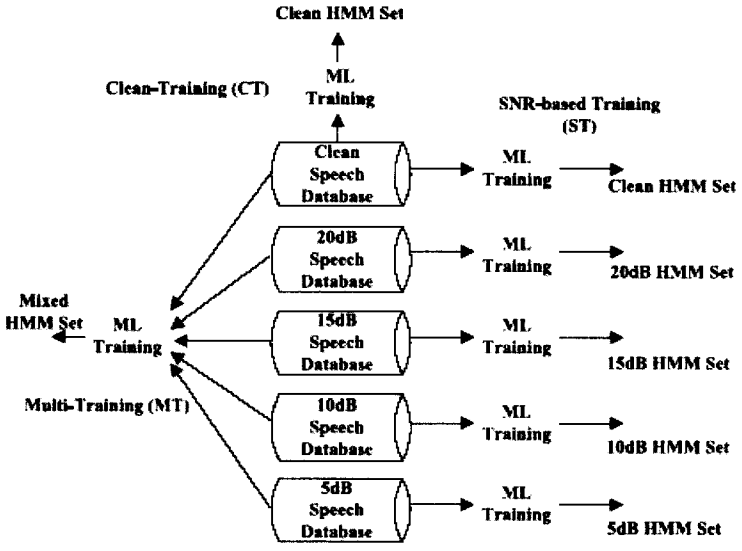


Figure 1. Illustration of three training modes

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 $\boldsymbol{W}_r$, we adopt KLD between the corresponding HMMs:

$$\mathcal{A}(\boldsymbol{W},\boldsymbol{W}_r) = -D(\boldsymbol{W}_r \,||\, \boldsymbol{W}) \tag{3}$$

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

Our goal is to measure the KLD for word sequences in Eq. 3. Given two word sequences $\boldsymbol{W}_r$ and $\boldsymbol{W}$ without their state segmentations, we should use a state matching algorithm to measure the KLD between the corresponding HMMs [Liu *et al.* 2005]. With state segmentations, the calculation can be further decomposed down to the state level:

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



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 assuming all observations are independent, we obtain:

$$D(\boldsymbol{s}_{\mathrm{r}}^{1:T} \,||\, \boldsymbol{s}^{1:T}) = \sum_{t=1}^{T} D(s_{\mathrm{r}}^{t} \,||\, s^{t}) = \sum_{t=1}^{T} \int p(\boldsymbol{o}^{t} \mid \boldsymbol{s}_{\mathrm{r}}^{t}) \log \frac{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}_{\mathrm{r}}^{t})}{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}^{t})} d\boldsymbol{o}^{t} \tag{5}$$

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

Now, our problem is how to measure the KLD between two states. Conventionally, each state s is characterized by a Gaussian Mixture Model (GMM): $p(\boldsymbol{o} \mid s) = \sum_{m=1}^{M_s} w_{sm} \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\boldsymbol{sm}}, \boldsymbol{\Sigma}_{\boldsymbol{sm}})$, 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 [Goldberger *et al.* 2003] has been proposed to approximate the KLD measurement of the two GMMs.

Let $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a N-dimensional Gaussian distribution and h be an arbitrary $\mathrm{IR}^{N} \rightarrow \mathrm{IR}$ function, the 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 $\lambda_k, \boldsymbol{u}_k$ are the k^{th} eigenvalue and eigenvector of $\boldsymbol{\Sigma}$, respectively. Geometrically, all these "*sigma*" points are on the principal axes of $\boldsymbol{\Sigma}$. Equation 6 is precise if h is quadratic.

For our case, the Gaussian distribution in Eq. 6 is replaced by a GMM, and the function h corresponds to the term $\log \frac{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}_{\mathrm{r}}^{t})}{p(\boldsymbol{o}^{t} \mid \boldsymbol{s}^{t})}$ in Eq. 5. Then, KLD between two states (GMMs) can be approximated by:

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

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

3.3 Gain Function Calculation

Usually, a word graph is a compact representation of large hypotheses space in speech recognition. As 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):

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

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

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

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

From the objective function defined in Eq. 1, the gain function $\mathcal{A}(W, W_r)$ is dependent on the model parameters, which should be updated in optimization process. In [Du *et al.* 2007], we conclude that the optimization of the gain function term has little impact on the performance. So here, $\mathcal{A}(W, W_r)$ is considered a constant term and not optimized. The KLDs related to gain function are precomputed using the ML trained model parameters. Then our optimization of objective function is the same as that mentioned in [Povey 2004]. We use the 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 TIDigits and Aurora2, both English 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 baseline configuration for two databases is listed in Table 2.

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	10	6	12549	12547
Aurora2			16	3	8440*2	1001*70

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 [Hirsch *et al.* 2000].

For minimum 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, we select the best results after 20 iterations for each sub set of testing.

4.2 Experiments on TIDigits Database

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

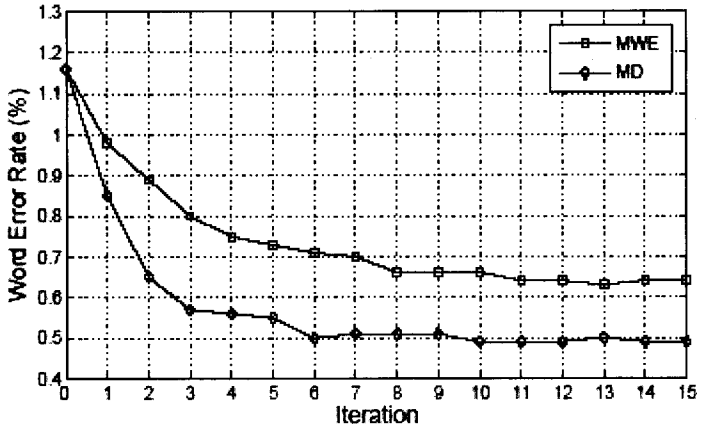


Figure 2. Performance comparison on TIDigits

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 the silence model in minimum error training using different training modes. Since it is unrelated to the criteria, here we adopt MWE. When applying clean-training, the performances of all test sets without updating silence model are consistently better. However, in multi-training, the conclusion is the opposite. From the results, we can conclude that increasing the discrimination of the silence model will lead to performance degradation in mismatched cases (clean-training) and performance improvement in matched cases (multi-training). This can be explained as follows: For the clean-training case, if we increase the discrimination of the silence model, the noise segments are more easily recognized as digits when testing on noisy data. Then, insertion errors will increase. However, for the multi-training case, the silence model represents both silence and noise segments, which is matched with that when testing on noisy data. So, by updating the silence model, the global performance will be improved. 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%
20dB	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%
15dB	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%
10dB	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%
5dB	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%
0dB	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%
20dB	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%
15dB	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%
10dB	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%
5dB	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%
0dB	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 is 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. This conclusion can be also drawn in clean-training and SNR-based training cases.

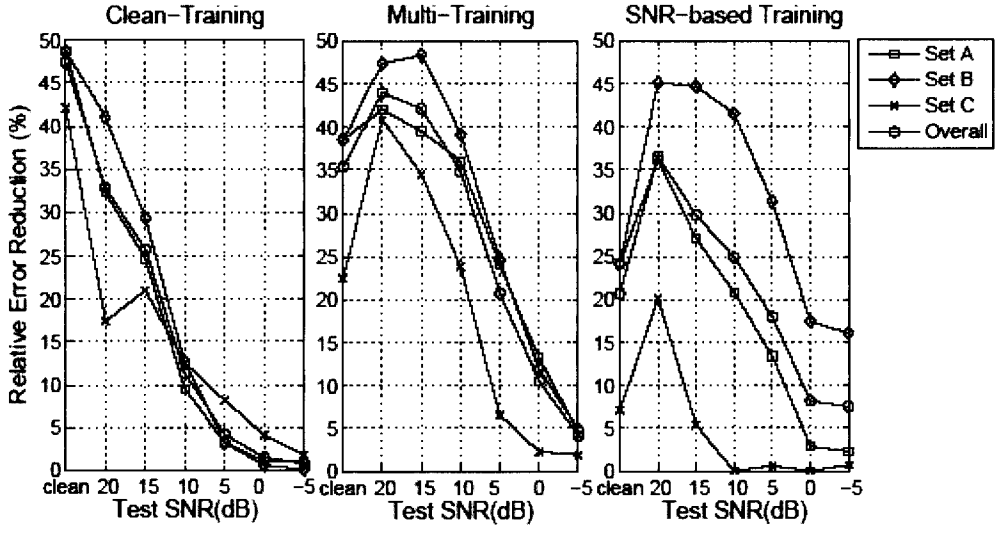


Figure 3. Relative Improvement over ML baseline on Aurora2 using different training modes in MD training

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%

Different Training Modes. Figure 3 shows relative improvement over ML baseline using MD training with different training modes. From this figure, some 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 is 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, was almost the worst in all training modes. Second, the relative improvement performance declines for decreasing SNR in clean-training. However, 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.

5. Conclusions

In this paper, the noise robustness of discriminatively trained HMMs is investigated. Discriminatively trained models are tested on English continuous digit databases, and MD and MWE criteria are experimentally compared to test the affection of error resolution. We observe: 1. Minimum error training is effective not only in clean environments, but also in noisy environments, which can be concluded in various training modes. Minimum error training is more effective as the SNR increases. Even when testing on mismatched noise scenarios, minimum error training also achieves better performance than ML training. 2. In minimum error training, higher resolution error analysis is more helpful at low SNRs. 3. Silence models should be carefully updated when the training and testing data are not well-matched.

Reference

Du, J., P. Liu, H. Jiang, F.K. Soong, and R.-H. Wang, "A New Minimum Divergence Approach to Discriminative Training," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2007, pp. 677-680.

Du, J., P. Liu, F.K. Soong, J.-L. Zhou, and R.-H. Wang, "Minimum Divergence Based Discriminative Training," In*Proceedings of International Conference on Spoken Language Processing*, 2006, pp. 2410-2413.

Gales, M.J.F. and S.J. Young, "Robust Continuous Speech Recognition using Parallel Model Combination," Technical Report EDICS Number: SA 1.6.8, Cambridge University, 1994.

Goldberger, J., "An Efficient Image Similarity Measure Based on Approximations of KL-Divergence between Two Gaussian Mixtures," In*Proceedings of International Conference on Computer Vision*, 2003, pp. 370-377.

Gong, Y., "Speech Recognition in Noisy Environments: A Survey," *Speech Communication*, 16, 1995, pp. 261-291.

Hirsch, H.G. and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions," In*Proceedings of ISCA ITRW ASR*, 2000, pp. 181-188.

Juang, B.-H., W. Chou, and C.-H. Lee, "Minimum Classification Error Rate Methods for Speech Recogtion," *IEEE Transactions on Speech and Audio Processing*, 5(3), 1997, pp. 257-265.

Kullback, S. and R.A. Leibler, "On Information and Sufficiency," *Ann. Math. Stat*, 22, 1951, pp. 79-86.

Laurila, K., M. Vasilache, and O. Viikki, "A Combination of Discriminative and Maximum Likelihood Techniques for Noise Robust Speech Recognition," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1998, pp. 85-88.

Liu, P., F.K. Soong, and J.-L. Zhou, "Effective Estimation of Kullback-Leibler Divergence between Speech Models," Technical Report, Microsoft Research Asia, 2005.

Meyer, C. and G. Rose, "Improved Noise Robustness by Corrective and Rival Training," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2001, pp. 293-296.

Ohkura, K., D. Rainton, and M. Sugiyama, "Noise-robust HMMs Based on Minimum Error Classification," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1993, pp. 75-78.

Povey, D., "Discriminative Training for Large Vocabulary Speech Recognition," PhD thesis, Cambridge University, 2004.

Schluter, R., "Investigations on Discriminative Training Criteria," PhD thesis, Aachen University, 2000.

Valtchev, V., J.J. Odell, P.C. Woodland, and S.J. Young, "MMIE Training of Large Vocabulary Speech Recognition Systems," *Speech Communication*, 22, 1997, pp. 303-314.

Varga, A.P. and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," In*Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 1990, pp. 845-848.

NATIONAL CENTRAL LIBRARY ROC