IMPROVED NOISE ROBUSTNESS BY CORRECTIVE AND RIVAL TRAINING

Carsten Meyer *Georg Rose*

Philips Research Laboratories
Weißhausstr. 2, D-52066 Aachen, Germany
e-mail: {Carsten.Meyer, Georg.Rose}@philips.com

ABSTRACT

We show that discriminative training methods have the potential to improve noise robustness even for high resolution acoustic models trained on noisy data. To this end, we compare the performance of acoustic models trained on noisy data using maximum likelihood (ML), corrective (CT) and rival training (RT). Experiments are performed on a German and a Dutch continuous digit string recognition task, yielding improvements in the range of 12% to 35% relative.

1. INTRODUCTION

In spite of considerable progress made in the past years, environmental noise is still one of the most challenging problems in practical applications of automatic speech recognition. Several strategies to handle this problem have been investigated: One possibility is try to eliminate the noise from the acoustic features (e.g. spectral subtraction [5], Wiener filtering or to enhance the speech component (e.g. singular value decomposition [3]). A second approach is to adapt the acoustic models to the noisy environment (e.g. MLLR and MAP [5], PMC [6]). In any case, however, it is advisable to choose a suitable training objective in acoustic model training which guarantees best performance in presence of noise. In this study, we investigate if discriminative training (DT) is suited to improve noise robustness as compared to conventional maximum likelihood (ML) training.

In various studies on clean data it has been shown that DT improves the performance of acoustic models as compared to ML training. Discriminative training criteria focus directly on misclassifications by increasing the class separability between the acoustic models. Such criteria include minimum classification error (MCE) [8], maximum mutual information (MMI) [1] and corrective training (CT) as special case of MMI training. Recently, an extended version of the corrective training algorithm, called "rival training" (RT), has been proposed [12]. As CT, this algorithm can be implemented completely within the Viterbi framework, but outperforms CT significantly.

Performing acoustic model training on *clean data*, DT has been applied successfully also to noise mismatch situations. For example, [11] has investigated the combination of minimum error classification and various algorithms for noise mismatched environments.

In many applications, however, the available training data contain noise, even after application of noise suppression methods. Furthermore, in order to decrease the mismatch between training and (noisy) test data one can simply add noise to the training data. Thus it is important to investigate the performance of discriminative training also on noisy training data. This scenario has been addressed, for example in [7], proposing a discriminative training algorithm for environmental parameters. In [10], it was found for "coarse" acoustic models (single densities and three densities per mixture) that MCE clearly outperforms ML in a noise robust speech recognition task (where the training utterances were recorded in three different noise conditions).

For *high resolution* acoustic models, however, it is generally known that the improvements gained by discriminative training are smaller than for low resolution models [14, 13]. The general problem is that the large number of parameters might lead to overfitting effects, i.e. performance degradation on independent test data. This effect might be even worse for training on noisy data, since discriminative training might adapt the models to the specific noise realisation, without improving the robustness of the acoustic models. Therefore, the goal of our work is to show that discriminative training in general has the potential for improving model quality even for large scale models trained on noisy data. For efficiency reasons, we are focussing on rival training and corrective training as the simplest discriminative training algorithms.

We compare in our work the performance of acoustic models with more than 60 densities per mixture, trained with ML, CT and RT on data characterized by additive noise, in a continuous digit string recognition task. Evaluation experiments are carried out in noise matched and mismatched conditions. For the noise mismatch situations, we do not assume any prior knowledge about the degree of noise of the test data. Experiments are reported on a German database with additive Gaussian white noise, and a Dutch database with additive real car noise, applying quite different feature extractions.

2. TRAINING CRITERIA

2.1. Maximum Likelihood Training

A commonly used training criterion to determine the parameters of the acoustic model is the maximum likelihood (ML) principle

$$F_{ML}(\lambda) = \sum_{r=1}^{R} \log p_\lambda(X_r|W_r) \ , \qquad (1)$$

where X_r denotes a sequence of acoustic observation vectors for utterance $r \in \{1, \ldots, R\}$, W_r denotes the corresponding sequence of spoken words and λ represents the set of acoustic model parameters. Maximum likelihood parameter estimation tries to maximize the likelihood for the spoken word sequence to generate the observed feature sequence. Since competing models are not taken into account, the recognition accuracy is maximized only indirectly.

2.2. Corrective and Rival Training

Several discriminative training (DT) approaches have been suggested [1, 2] trying to maximize class separability and thus to improve recognition accuracy directly. A commonly used approach is to maximize

$$\begin{aligned} F_{MMI}(\lambda) &= \sum_{r=1}^{R} \log p_\lambda(W_r|X_r) \\ &= \sum_{r=1}^{R} \log \frac{p_\lambda(X_r|W_r)\,p(W_r)}{p_\lambda(X_r|W_{gen})} \qquad (2) \end{aligned}$$

with

$$p_\lambda(X_r|W_{gen}) := \sum_{W} p_\lambda(X_r|W)\,p(W) \ . \qquad (3)$$

This training criterion ("maximum mutual information", MMI) simultaneously tries to increase the likelihood of the spoken word sequence W_r and to decrease the likelihood of competing hypotheses W contained in the "general model" W_{gen}. In general, W_{gen} is obtained by a recognition pass on the training data. The acoustic parameters λ are then reestimated in an iterative procedure, involving the determination of W_{gen} according to the new estimate of λ and the re-application of the reestimation equations in each iteration step. The reestimation equations are given e.g. in [13].

The implementation of this training process can be greatly simplified if the general model (3) is restricted to the *recognized text*. Since in this case correctly recognized sentences cancel out in equation (2), only misrecognized sentences contribute to the training criterion. The resulting algorithm is called *corrective training* (CT).

Corrective training, however, has its limits when the training error is very low, since then only few material is used for reestimation of the acoustic parameters, with increasing risk of overfitting. A simple extension of the CT algorithm, called "rival training" (RT) has been proposed in [12], which is implementationally much less expensive than lattice–based discriminative training methods like MMI and MCE, but gives significantly better performance as CT. The algorithm uses also correctly recognized sentences, if their (absolute) score difference to the second best hypothesis does not exceed a threshold value, determined as quantil of the score difference histogram [12]. This amounts to defining W_{gen} to be the best scored *incorrect* hypothesis.

3. EXPERIMENTS

3.1. Noise addition

In our experiments, we use the following noise addition scheme: For each utterance of the (clean) data, the SNR of the utterance is estimated. Then, given a "target SNR" (i.e. the desired SNR of each utterance after noise addition), a weight for noise addition is calculated for each utterance such that the (weighted) addition of random noise to the clean signal yields the desired target SNR. The noise addition is performed on the sample level.

For noise addition, we use two scenarios: Gaussian white noise and real car noise. The car noise was taken from the MoTiV [9] data collection, using data recorded in 5 cars, each at two different speed values (city and highway). For each utterance, the noise environment to be added is chosen at random.

3.2. Training scheme

Experiments were performed on a continuous digit string recognition task on a German and a Dutch database. We use continuous Gaussian mixture emission distributions with a globally pooled variance vector.

The general training scheme proceeds as follows: First, we apply ML training using Viterbi alignment and the maximum approximation. Then, using the ML references as baseline, CT and RT are performed by iterating the reestimation equations given in [13]. In CT, in each iteration step we use the recognized text, obtained by a recognition pass on the training corpus, as the general model W_{gen}. In RT, in each iteration step we determine the "rivalizing text" from a 2–Best–List (obtained by a recognition pass) and the spoken text according to the algorithm presented in [12]. After a predefined number of iterations of CT or RT, respectively, we select the references yielding the lowest error rate on the training corpus for evaluation.

3.3. Experiments on German digits

In the first series of experiments, we used the male part of the *SieTill* corpus [4] for telephone line recorded German connected digit strings. This corpus consists of about 23k spoken digits in 7k sentences (190 speakers, about 2.5h of speech) for both training and test. The acoustic model consists of whole word HMMs (11 models including "*zwo*" as synonyme for the digit "2") plus one silence model. Experiments are reported with high resolution acoustic models (28k densities, 108 per mixture). We used 11 cepstral features plus derivatives of the first 9 coefficients. The sampling rate was 8kHz, the frame shift 16ms. We applied cepstral mean subtraction on the sentence level and a LDA transform resulting in a 24 component feature vector. Noise addition was performed as explained in section 3.1 at 12dB SNR. Figure 1 shows the word error rates (WER) on the training corpus in dependence of the iteration number.

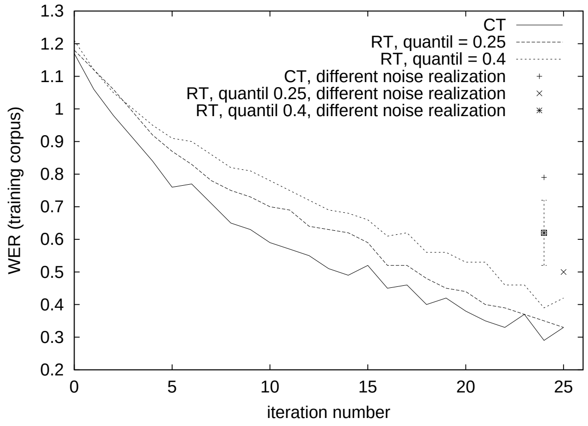

Figure 1: Word error rates (WER) in dependence of the iteration number for corrective (CT) and rival training (RT) on the SieTill male training corpus with additive Gaussian white noise, target SNR of 12dB. The isolated points ($+, \times, *$) represent the WER for a different noise realisation (at the same target SNR of 12dB) for the respective iteration number. The errorbar indicates the 95% confidence level.

In general the individual recognition passes involved in each iteration of discriminative training are performed with the given *fixed* noise realisation. Using artificial noise, one has the opportunity to evaluate the estimated acoustic models for *different* noise realisations, at the same target SNR. In these lines, we have estimated discriminativeley trained acoustic models using a fixed noise realisation, and then evaluated them again on the training corpus using a different noise realisation. These experiments indicate to what extent the discriminatively trained models have "adapted" to the specific noise realisation, instead of improving model quality. Indeed, such a fitting to the noise realisation has been observed, see figure 1, symbols $+, \times, *$. Nevertheless, the evaluation results are significantly better for the discriminatively trained models than the corresponding result for the ML baseline (iteration number 0 in figure 1). It is particularly interesting to observe less noise fitting for the RT models than for CT.

Evaluation results on the SieTill male test corpus are presented in table 1, for maximum likelihood (ML) models, corrective training (CT) and rival training (RT). The first table a) shows recognition results on test data with additive Gaussian white noise, target SNR of 12dB, corresponding to noise matched conditions. The second table b) presents recognition results on clean test data, i.e. noise mismatched conditions. The word penalty used for the evaluation experiments has been optimized on the training corpus, separately for the ML, CT and RT models.

a) test data with target SNR of 12dB						
	qu.	it.	WER	it.	WER	rel. impr.
ML	baseline error rate:(3.86 ± 0.26)%					
CT	0.0	16	3.63	24	3.61	-6.5%
RT	0.25	16	3.51	25	3.40	-11.9%
RT	0.4	16	3.58	24	3.40	-11.9%

b) clean test data						
	qu.	it.	WER	it.	WER	rel. impr.
ML	baseline error rate:(4.61 ± 0.28)%					
CT	0.0	16	4.34	24	4.37	-5.2%
RT	0.25	16	4.27	25	4.03	-12.6%
RT	0.4	16	4.08	24	3.84	-16.7%

Table 1: Evaluation results on the SieTill male test corpus, a) with additive Gaussian white noise, target SNR of 12dB (i.e. noise matched conditions), b) on clean test data (noise mismatched conditions). The acoustic models were trained on SieTill training data with additive Gaussian white noise with target SNR of 12dB, according to maximum likelihood (ML), corrective (CT) and rival training (RT). "it." denotes the number of discriminative training iterations, "qu." the quantil value, and "rel. impr." the relative improvement compared to the ML baseline result.

Comparing the CT results after 16 and 24 iterations, we obtained similar error rates on the test data, although the training error rate still decreases (figure 1). In contrast, RT yields still improvements on the test corpus continuing training up to 25 iterations.

Summarizing, it can be seen that the discriminatively trained models outperform the ML models on both noise matched and noise mismatched conditions. The improvements gained by rival training are larger than those obtained by corrective training, and are significant on the 95% confidence level.

3.4. Experiments on Dutch digits

The second part of our experiments is carried out on the *Polyphone* corpus for telephone line recorded Dutch connected digit strings. For training, we used 4k sentences of the Polyphone training corpus (31k digits, about 4.5h of speech). The test corpus consists of 2k sentences (15k digits). We used whole word HMMs (11 models including a pronunciation variant for the digit "7") plus one silence model, in total 19k densities, 63 per mixture. A different feature extraction was performed: here, we used 14 spectral coefficients and their first derivatives plus the energy with first and second derivative, and applied recursive long-term spectrum normalization (sampling rate and frame shift as in section 3.3). Applying a LDA transform resulted in a 31 component feature vector. The addition of car noise was performed as described in section 3.1.

Evaluation experiments on the Polyphone test corpus are carried out for noise matched conditions (test data with additive car noise of target SNR of 12dB) and mismatched conditions (additive car noise with target SNR of 6dB and clean test data). Table 2 presents evaluation results for maximum likelihood (ML) and rival training (RT, 12 iterations, quantil 0.4). These experiments are carried out with a *fixed* word penalty of 90000, which was optimized on the training corpus for the ML models.

noise level		WER	rel. gain
12dB test data	ML	8.37 ± 0.5	—
("matched")	RT	5.66 ± 0.4	-32.4%
6dB test data	ML	18.77 ± 0.6	—
("mismatched")	RT	12.16 ± 0.5	-35.2%
clean test data	ML	11.17 ± 0.5	—
("mismatched")	RT	9.44 ± 0.5	-15.5%

Table 2: Evaluation results on the Polyphone test corpus with additive car noise of target SNR of 12dB (noise matched conditions), target SNR of 6dB and clean test data (noise mismatched conditions). Maximum likelihood (ML) and rival training (RT, 12 iterations) was performed on Polyphone training data with additive car noise, target SNR of 12dB.

For all investigated conditions (matched: target SNR of 12dB in both training and test, mismatched: different target SNR in test and in training), we obsered significant performance gains (15% up to 35%) by RT as compared to ML.

4. DISCUSSION

We investigated the performance of maximum likelihood, corrective and rival training for large scale acoustic models trained on data with additive noise. The experiments demonstrated that in spite of fitting the models to the specific noise realisation, corrective and rival training outperform maximum likelihood training significantly. We conclude that discriminative training methods have the potential to improve noise robustness even for high resolution acoustic models trained on noisy data.

In general, no prior information about the noise level of the test data is available. This requires to use a fixed word penalty, which was in our experiments tuned on the training data. In further experiments we observed that in noise mismatched conditions, the optimal word penalty depends on the noise level of the test data. Optimizing the word penalty "a posteriori" on the *test* data, we found that the RT acoustic models are less sensitive on that parameter than the ML models. However, in some mismatched conditions, the performance of the "optimal" ML models (with respect to the word penalty) was slightly better than for RT.

In future work, it should be investigated if the noise robustness gained by discriminative training will add to the improvements obtained by standard noise robustness techniques. Also, a comparison with lattice based discriminative training methods (MMI, MCE) might be performed.

Acknowledgement: It is a pleasure to thank Peter Beyerlein and Alexander Fischer for valuable comments.

5. REFERENCES

1. Bahl, L. R., Brown, P. F., de Souza, P. V. and Mercer, R. L., "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", *Proc. ICASSP-86*, pp. 49-52, Tokyo, 1986
2. Chou, W., Lee, C.-H. and Juang, B.-H, "Minimum Error Rate Training based on N-Best String Models", *Proc. ICASSP-93*, pp. 652-655, Minneapolis, MN, 1993
3. De Moor, B., "The singular value decomposition and long and short spaces of noisy matrices", *IEEE Trans. SP* **41**: 2826-2838, 1993
4. Eisele, T., Haeb-Umbach, R. and Langmann, D., "A comparative study of linear feature transformation techniques for automatic speech recognition", *Proc. ICSLP-96*, pp. 252-255, Philadelphia, PA, 1996
5. Fischer, A. and Stahl, V., "Database and Online Adaptation for Improved Speech Recognition in Car Environments", *Proc. ICASSP-99*, Vol. 1, pp. 445-449, Phoenix, USA, 1999
6. Gales, M. J. F., "Model–Based Technique for Noise Robust Speech Recognition", *Dissertation, University of Cambridge*, 1995
7. Han, J. *et al*, "Discriminative learning of additive noise and channel distortions for robust speech recognition", Proc. *Int. Conf. Acoustics, Speech and Signal Process. 1998*, Seattle, WA, Vol. 1, pp. 81-84, May 1998
8. Juang, B. H. and Katagiri, S., "Discriminative learning for minimum error classification", *IEEE Trans. SP* **40**: 3043-3054, 1992
9. Langmann, D., Schneider, T., Grudszus, R., Fischer, A., Crull, T., Pfitzinger, H., Westphal, M. and Jekosch, U., "CSDC — The MoTiV Car–Speech Data Collection", in *First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998
10. Laurila, K., Vasilache, M., and Viikki, O. "A combination of discriminative and maximum likelihood techniques for noise robust speech recognition", Proc. *Int. Conf. Acoustics, Speech and Signal Process. 1998*, Seattle, WA, Vol. 1, pp. 85-88, May 1998
11. Lin, M. T., Spanias, A. and Loizou, P., "An improved approach to robust speech recognition using minimum error classification", *Speech Communication* **30**:27-36, 2000
12. Meyer, C. and Rose, G., "Rival Training: Efficient Use of Data in Discriminative Training, *Proc. ICSLP-00*, Vol. IV, pp. 632 - 635, Beijing, 2000
13. Schlüter, R., and Macherey, W: "Comparison of discriminative training criteria", *Proc. ICASSP-98*, pp. 493-496, Seattle, WA, 1998
14. Woodland, P. C., and Povey, D., "Large Scale Discriminative Training for Speech Recognition", *Proc. ASR-2000*, pp. 7-16, Paris 2000