ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Improving Discriminative Training for Robust Acoustic Models in Large Vocabulary Continuous Speech Recognition

Janne Pylkkönen and Mikko Kurimo

Department of Information and Computer Science, Aalto University, Finland
janne.pylkkonen@aalto.fi, mikko.kurimo@aalto.fi

Abstract

This paper studies the robustness of discriminatively trained acoustic models for large vocabulary continuous speech recognition. Popular discriminative criteria maximum mutual information (MMI), minimum phone error (MPE), and minimum phone frame error (MPFE), are used in the experiments, which include realistic mismatched conditions from Finnish Speecon corpus and English Wall Street Journal corpus. A simple regularization method for discriminative training is proposed and it is shown to improve the robustness of acoustic models gaining consistent improvements in noisy conditions.

Index Terms: speech recognition, discriminative training, robustness

1. Introduction

In real-life speech recognition applications it is often difficult to have sufficient control over the recognition conditions. Robustness to e.g. different background noises can therefore be a crucial aspect in the system design. Still most speech recognition techniques are validated only in well-matching settings. However, using noisy test sets and out-of-domain evaluation data may present completely different behavior than the clean and matching conditions provide.

Discriminative training is nowadays an established way for improving acoustic models in speech recognition. With proper training heuristics discriminatively trained acoustic models provide significant improvements over the traditional maximum likelihood (ML) training. Yet surprisingly few studies have been conducted on how discriminative training affects the robustness of acoustic models and whether the established training methods are useful when considering mismatched recognition conditions. MMI training has been shown to be beneficial in cross-task conditions [1]. More recently, the generalization ability of discriminatively trained acoustic models were tested in noisy digit recognition task [2], studying soft-margin estimation (SME) and minimum classification error (MCE). These studies show that discriminative training can have an important role in noise robust speech recognition.

This paper studies the robustness of discriminatively trained acoustic models. Three well-known discriminative criteria are experimented, namely maximum mutual information (MMI), minimum phone error (MPE), and minimum phone frame error (MPFE). The acoustic models for the study are trained on clean speech and their performance are evaluated in matching clean and various mismatched noisy conditions. A new discriminative training regularization method is proposed and it is shown to improve the robustness of the acoustic models. This study concentrates on the effect of discriminative training to the robustness of the acoustic models, so feature-domain and other model-domain techniques for improving noise robustness (e.g. [3, 4]) have been omitted.

2. Discriminative Training

2.1. MMI

The earliest successful discriminative training method for LVCSR acoustic models was the maximum mutual information (MMI) estimation [5]. MMI can be seen as a direct extension of the ML estimation where the training criterion, the likelihood of the training data, is replaced with the conditional likelihood. For the training procedure this simple change has dramatic consequences. In order to compute the conditional likelihood, MMI estimation needs to consider all the possible recognition hypotheses of the training utterances, creating substantial computational challenges. An efficient way to deal with this is to represent the alternative hypotheses in a form of a lattice. The parameter estimation for MMI and other discriminative criteria is best done with the extended Baum-Welch (EBW) [5] algorithm. Also several training heuristics are needed in order to make MMI estimation feasible for LVCSR [5].

2.2. MPE and MPFE

Although MMI improves the discriminative capability of acoustic models compared to ones trained with ML, a discriminative criterion can be even more explicit in its goal. A natural goal for acoustic model training is the minimization of recognition errors. Povey and Woodland [6] introduced the minimum word error (MWE) and minimum phone error (MPE) criteria, of which the latter turned out to be better in minimizing the error rates of an independent test set. In MPE the training criterion is an approximation of the expected phone error over the training set.

Several alternative formulations for an error minimizing criterion have been proposed since the introduction of MPE. One of the most successful one, minimum phone frame error (MPFE) [7], replaces the phone level error approximation with a simple frame level error. Both MPE and MPFE were used in the experiments of this study. MPFE was implemented with the modifications suggested in [8]. Despite the differences in the training criteria, both MPE and MPFE training can use the same parameter estimation methods applied with MMI.

2.3. Controlling EBW

When performing discriminative training with the EBW algorithm, a crucial smoothing constant D needs to be set such that proper convergence of the discriminative criterion is achieved. Woodland and Povey [5] suggested the widely accepted heuristic method for setting different D constants for each Gaussian in the acoustic model. It sets the EBW constant D_i of Gaussian i as the maximum of:

1. Denominator occupancy multiplied by two ($2\gamma_i^{den}$)
2. Double the value necessary for the Gaussian covariance to be positive definite ($2D_i^{\min}$)

The denominator occupancy of a Gaussian γ_i^{den} is the sum of probability masses of that Gaussian in the recognition model, summed over all the training utterances. In practice, the first case defines the constant D_i for majority of Gaussians.

With the introduction of the MPE criterion, a need for additional regularization for EBW arose. The heuristics developed for MMI resulted in poor generalization ability with the MPE criterion and did not improve the recognition accuracy on an independent test set. As a remedy, a smoothing method for discriminative statistics called I-smoothing [6] was proposed. It provided the required additional regularization for the training to succeed. The proper I-smoothing value is criterion dependent, and for example MPFE requires larger smoothing values than the original MPE [8]. Later Povey *et al.* proposed an alternative formulation for I-smoothing which is equivalent to adding a constant value to the Gaussian specific D_i [9].

If I-smoothing is defined as an addition of a constant to D_i, the effect of denominator occupancies to the Gaussian specific EBW constants is diminished. This observation motivated us to test discriminative training with a global D that is the same for all the Gaussians. To ensure proper parameter estimation, the case 2 in the above heuristics was still maintained, although it is typically applied with only a small fraction of the Gaussians. The new method for setting the Gaussian specific EBW constant can then be expressed as

$$D_i = \max\{D^{\text{global}}, 2D_i^{\min}\}. \qquad (1)$$

Preliminary experiments showed this method of setting the EBW constants to perform similarly as the usual heuristics when tested on matching clean speech. However, noisy tasks showed consistent improvements over the traditional method. With a global D, EBW applies similar regularization to (almost) all Gaussians, instead of heuristically adjusted regularizations with Gaussian specific D_i.

For setting the global D to control the EBW algorithm, we propose a new simple method. In an effort to obtain interpretable and criterion independent method for selecting this crucial constant, we relied on measuring the Kullback-Leibler divergence (KLD) [10] between the Gaussians. The global D is sought such that the first model update of the discriminative training obtains a desired level of model change, as measured by the median of the KLDs between the original and updated Gaussians. The computations are simple as the KLD between the Gaussians can be computed with a closed-form formula:

$$\begin{aligned}&\mathcal{D}\big(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \,\|\, \mathcal{N}(\boldsymbol{\mu}_i^0, \boldsymbol{\Sigma}_i^0)\big)\\ &= \frac{1}{2}\Big[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^0)^T(\boldsymbol{\Sigma}_i^0)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_i^0)\\ &\qquad + \operatorname{tr}\left[(\boldsymbol{\Sigma}_i^0)^{-1}\boldsymbol{\Sigma}_i\right] + \log\frac{|\boldsymbol{\Sigma}_i^0|}{|\boldsymbol{\Sigma}_i|} - d\Big] \qquad (2)\end{aligned}$$

where $\boldsymbol{\mu}_i^0$ and $\boldsymbol{\Sigma}_i^0$ are the mean and covariance of the original Gaussian, and $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance of the updated Gaussian, respectively. d is the dimensionality of the models. Furthermore, the iterative searching of the global D only requires the discriminative statistics of the training data as computed with the initial model. The additional computations required are therefore marginal.

Table 1: Finnish Speecon development and evaluation sets.

	Development			Evaluation		
Condition	#spkr	#utt	length	#spkr	#utt	length
Clean/Ch2	40	1093	1.9h	40	1118	1.9h
Car	10	288	0.5h	20	575	1.0h

3. Experiment setup

Speech recognition experiments were conducted to assess the benefits of the global D approach for controlling EBW as well as to study the KLD based control strategy with several discriminative criteria. The LVCSR experiments were performed with Finnish Speecon corpus [11] and the English Wall Street Journal (WSJ) corpus [12] to demonstrate recognition performances in different conditions. The acoustic models were trained using only clean data. However, both clean and various mismatched noisy test sets were used for evaluation to assess the robustness of the discriminatively trained acoustic models.

The experiments were run with the speech recognition system developed at the Aalto University [13]. The acoustic models used three-state Hidden Markov models with Gaussian mixture emission probabilities. The number of diagonal Gaussians in the mixtures varied depending on the amount of data assigned for each state. The acoustic features were standard MFCCs with first and second differential features, 39 dimensions in total. ML models were first trained for initial models and to create lattices with unigram language models for the discriminative training. Discriminative training consisted of 15 iterations.

The training set for the Finnish acoustic models was extracted from the Finnish Speecon corpus. Only clean sentences were used for training. The training set consisted of 310 speakers and about 15h of speech in total. The acoustic models had 24587 Gaussians in 1170 mixtures/states. Table 1 summarizes the test sets extracted from the same Speecon corpus. In addition to the clean development and evaluation sets, two mismatched conditions were utilized. The "Clean Ch2" sets reused the clean recording sessions, but a medium distance (channel 2) microphone was used instead of the close talking one. Not all the environments in that set were quiet, so in addition to small reverberation effects also several types of background noises were present. In addition to this, a separate test set featuring car noise was used. That set was recorded in moving cars with a lavalier microphone. Except for the "Clean" and "Clean/Ch2" test sets, the speakers in all the different sets were disjoint.

The language model for the Finnish recognition experiments was a high-order morph-based N-gram model [13]. As the number of morphs that constitute a word is not limited, the resulting vocabulary is unlimited. Finnish words are commonly rather long and morphologically complex, so instead of word error rate (WER) letter error rate (LER) was used to improve the resolution of the error measurements.

The acoustic models for the English experiments were trained with the WSJ1 corpus which consists of 284 speakers and about 80h of speech. The models had 113227 Gaussians in 3569 mixtures/states. For evaluating clean recognition two sets were used. The actual evaluation was performed with the official WSJ 20k-word vocabulary evaluation set with 8 speakers and 333 sentences. The 10-speaker development set of Nov'93 H1P0 task [14] with 503 utterances was used for development purposes. The mismatched evaluations were conducted with two WSJ spoke tests, S7 and S8 [14], which present real recordings in various noisy environments. These tests used the record-

Table 2: *Finnish Speecon development set results (LER). The discriminative training with EBW was controlled either with a baseline heuristics and I-smoothing or a global D set to produce different median Gaussian KLD changes.*

Model	Clean	Ch2	Car
ML	3.0%	10.1%	38.1%
MMI			
I-smooth 0	2.9%	9.3%	32.6%
I-smooth 400	2.9%	8.6%	29.2%
I-smooth 800	2.9%	8.5%	28.9%
KLD 0.1	**2.8%**	10.1%	32.4%
KLD 0.02	**2.8%**	8.7%	29.1%
KLD 0.002	**2.8%**	7.8%	**25.2%**
KLD 0.0004	2.9%	**7.7%**	25.8%
MPE			
I-smooth 100	**2.6%**	9.1%	36.3%
I-smooth 400	**2.6%**	9.0%	36.0%
I-smooth 800	**2.6%**	9.1%	36.0%
KLD 0.4	2.8%	9.0%	**33.6%**
KLD 0.1	2.7%	**8.8%**	34.3%
KLD 0.02	2.7%	**8.8%**	35.8%
KLD 0.002	**2.6%**	8.9%	35.8%
MPFE			
I-smooth 100	**2.7%**	9.4%	38.1%
I-smooth 400	**2.7%**	9.2%	37.6%
I-smooth 800	**2.7%**	9.2%	37.6%
KLD 0.4	2.9%	8.6%	32.2%
KLD 0.1	**2.7%**	**8.1%**	**30.8%**
KLD 0.02	**2.7%**	8.5%	33.5%
KLD 0.002	**2.7%**	8.8%	35.5%

Table 3: *Finnish Speecon evaluation set results (LER).*

Model	Clean	Ch2	Car
ML	3.3%	10.0%	31.2%
MMI I-smoothing	3.2%	8.6%	23.4%
MMI global D	**3.0%**	**7.7%**	**20.4%**
MPE I-smoothing	**2.9%**	9.0%	29.4%
MPE global D	3.0%	**8.8%**	**27.6%**
MPFE I-smoothing	**2.9%**	9.2%	30.2%
MPFE global D	**2.9%**	**8.2%**	**23.9%**

Table 4: *WSJ development set results in (WER).*

Model	Clean	Noisy S7	Noisy S8
ML	14.6%	24.9%	14.9%
MMI			
I-smooth 0	13.6%	24.8%	14.1%
I-smooth 800	13.5%	24.6%	13.9%
KLD 0.1	13.5%	**23.7%**	13.2%
KLD 0.02	13.5%	23.8%	**12.9%**
KLD 0.002	**13.4%**	25.0%	**12.9%**
MPE			
I-smooth 50	13.2%	24.1%	13.2%
I-smooth 400	**13.0%**	23.6%	13.4%
KLD 0.1	13.3%	**22.6%**	**13.0%**
KLD 0.02	13.2%	22.8%	13.2%
KLD 0.002	13.1%	22.9%	13.5%
MPFE			
I-smooth 400	13.3%	22.2%	13.8%
KLD 0.1	13.5%	**19.8%**	**13.3%**
KLD 0.02	13.5%	21.2%	13.5%
KLD 0.002	**13.2%**	22.2%	13.5%

ings from the stand-mounted microphone. Only SNR 10dB conditions were used for S8. Both spoke tests had separate development and evaluation sets, each having 10 speakers and about 200 utterances. No adaptation or noise compensation was used for the acoustic models (unlike in the results reported in [14]). The noisy experiments used the 5k-word vocabulary for recognition. All the English recognition experiments used the 3-gram language models provided with the WSJ corpus. The recognition accuracy was evaluated with WER.

The baseline method for discriminative training was EBW using the common heuristics discussed in Section 2.3. I-smoothing to the previous model was used. For each evaluation set, the corresponding development set was used to pick the best model among the 15 discriminative training iterations. Also the best values for the global D and I-smoothing was chosen based on the development set results.

4. Results

Table 2 shows the development set results for the Speecon corpus, obtained by selecting the best performing model for each task among the 15 discriminative training iterations. The best results for each criterion and test set combination are bolded. The global D method shows similar performance with the baseline in the clean cases, but outperforms the baseline heuristics in the mismatched tasks. Analyzing the development set results of different training iterations revealed that the optimal number of discriminative training iterations was very different among the different sets without any evident pattern. Using a development set to select the best model was especially important for MMI with a global D, for which the recognition errors exhibited some oscillation in the later iterations. Almost all the tested global D values worked well with the clean recognition test, but for noisy tasks it was more important to select the optimal value. MPE and MPFE showed benefits with rather large update steps, but for MMI substantially smaller update steps were optimal.

The evaluation set results for the Speecon corpus are shown in Table 3. The clean ML result matches with the best previous results [15]. The evaluation set results confirm the development set results that using a global D in EBW instead of the baseline heuristics gives consistent improvements in mismatched noisy tasks with all the discriminative criteria. These improvements were statistically significant according to the Wilcoxon signed rank test ($\alpha = 5\%$) in all the mismatched evaluations. Somewhat surprisingly, the MMI criterion provided the best performing models for mismatched recognition.

Table 4 shows the development set results for the WSJ corpus, similar to Table 2. Again the global D method gives better results than the baseline method. The biggest differences compared to the Speecon results are with the MMI criterion. It no longer shows significant improvements over the other discriminative criteria in the noisy tasks. The optimal KLD values for global D with MMI criterion were also somewhat different. With WSJ, all the criteria gave better results with small update steps when tested with clean data, whereas bigger update steps were beneficial for the noisy tasks.

The evaluation set results for the WSJ corpus are shown in Table 5. The clean results indicate reasonable baseline for the comparison, although they are not the all-time best for the task. The evaluation set results are in line with the development set

Table 5: *WSJ evaluation set results (WER).*

Model	Clean	Noisy S7	Noisy S8
ML	10.4%	22.9%	22.3%
MMI I-smoothing	**9.7%**	23.1%	19.9%
MMI global D	9.9%	**21.9%**	**19.5%**
MPE I-smoothing	9.4%	21.6%	**18.8%**
MPE global D	**9.3%**	**20.5%**	**18.8%**
MPFE I-smoothing	**9.4%**	19.4%	19.5%
MPFE global D	**9.4%**	**17.7%**	**19.1%**

ones, showing that a global D method performs equally to or better than the baseline EBW with all the discriminative criteria. The improvements observed in S7 set were statistically significant according to the Wilcoxon signed rank test ($\alpha = 5\%$).

5. Discussion and conclusion

The recognition experiments exposed the clean acoustic models to multiple sources of mismatches: reverberation, different microphones, and background noises such as music and car noise. The acoustic features were basic MFCC features with only a simple cepstral mean subtraction channel compensation method. It is therefore encouraging that with proper settings, discriminative methods were able to improve the recognition results in all the different tasks with realistic noisy conditions.

The Gaussian specific smoothing constants and I-smoothing are the prevailing method for controlling the EBW algorithm. The conducted experiments showed clear evidence that it is instead beneficial to use a global D in EBW when aiming for robust acoustic models. In almost all mismatched cases the global D method performed better than the baseline method. The clean results showed similar accuracies as the baseline. The proposed method for setting the global D is computationally very feasible, as the tuning is done only for the first discriminative update. The tuning, on the other hand, is possible using the statistics alone, without need to iterate over the training data. The optimal initial model change varied among corpora, criteria and tasks, so it is a good idea to try different values for any particular case. However, most combinations resulted in improvements compared to the baseline EBW.

It is generally agreed that MPE and MPFE criteria can produce better models than the MMI criterion. It is therefore interesting that in mismatched conditions in the Speecon corpus, MMI consistently outperformed other criteria with clear margins. However, in WSJ mismatched evaluations, MMI was performing slightly worse than MPE/MPFE. It is not clear why such performance variations did occur. These results suggest that using MMI and MPFE as complementary models in recognition result combination can improve the robustness of the system even further.

Running several recognition tasks with the same acoustic models demonstrated that the optimal number of discriminative training iterations varies between the tasks. In many cases the recognition accuracy started degrading or even oscillating after a certain training iteration. However, no clear pattern on this behavior was observed. Observing the discriminative criterion alone is insufficient in controlling the number of discriminative training iterations even with the matched tasks. It is therefore important to determine the proper number of iterations with a development set which is as realistic as possible.

Robustness of the discriminatively trained acoustic models has not received the attention it deserves considering the widely spread use of discriminative training and importance of robustness to real-life recognition tasks. The simple regularization method for EBW presented in this paper demonstrated clear improvements to the robustness of the acoustic models. Future work includes exploring the interaction of discriminative training with noise robust front-ends and noise compensation methods and evaluating robustness of discriminative methods when performing multicondition training.

6. Acknowledgements

This work was supported by the Academy of Finland in the projects numbered 135003, 140969 and 251170, and by Tekes in Perso project.

7. References

[1] R. Cordoba, P. Woodland, and M. Gales, “Improved cross-task recognition using MMIE training,” in *Proc. ICASSP*, 2002, pp. 85–88.

[2] X. Xiao, J. Li, E. S. Chng, H. Li, and C.-H. Lee, “A study on the generalization capability of acoustic models for robust speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1158–1169, 2010.

[3] A. Sankar and C.-H. Lee, “A maximum-likelihood approach to stochastic matching for robust speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 4, no. 3, pp. 190–202, 1996.

[4] 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 Audio Process.*, vol. 12, no. 3, pp. 218–233, 2004.

[5] P. C. Woodland and D. Povey, “Large scale discriminative training of hidden Markov models for speech recognition,” *Comp. Speech and Lang.*, vol. 16, pp. 25–47, 2002.

[6] D. Povey and P. C. Woodland, “Minimum phone error and I-smoothing for improved discriminative training,” in *Proc. ICASSP*, 2002, pp. 105–108.

[7] J. Zheng and A. Stolcke, “Improved discriminative training using phone lattices,” in *Proc. Interspeech*, 2005, pp. 2125–2128.

[8] D. Povey and B. Kingsbury, “Evaluation of proposed modifications to MPE for large scale discriminative training,” in *Proc. ICASSP*, 2007, pp. 321–324.

[9] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, “Boosted MMI for model and feature-space discriminative training,” in *Proc. ICASSP*, 2008, pp. 4057–4060.

[10] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, 2005.

[11] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, “SPEECON - speech databases for consumer devices: Database specification and validation,” in *Proc. LREC*, 2002, pp. 329–333.

[12] D. B. Paul and J. M. Baker, “The design for the wall street journal-based CSR corpus,” in *HLT ’91: Proceedings of the workshop on Speech and Natural Language*, 1992, pp. 357–362.

[13] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, “Importance of high-order N-gram models in morph-based speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 724–732, 2009.

[14] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, “1993 benchmark tests for the ARPA spoken language program,” in *Proceedings of the workshop on Human Language Technology*, ser. HLT ’94, 1994, pp. 49–74.

[15] J. Pylkkönen, “Investigations on discriminative training in large scale acoustic model estimation,” in *Proc. Interspeech*, 2009, pp. 220–223.