Discrimininative training of narrow band - wide band adapted systems for meeting recognition

Martin Karafiát[1], *Lukáš Burget*[1], *Thomas Hain*[2], *and Jan Černocký*[1]

[1]Brno University of Technology, Speech@FIT, Faculty of Information Technology, Czech Republic
[2]University of Sheffield, Department of Computer Science, Sheffield S1 4DP, UK
{karafiat,burget,cernocky}@fit.vutbr.cz,t.hain@dcs.shef.ac.uk

Abstract

The amount of training data has a crucial effect on the accuracy of HMM based meeting recognition systems. One of the largest collections of speech data is conversational telephone speech which was found to match speech in meetings well. However it is naturally recorded with limited bandwidth. In previous work we presented a scheme that allows to transform wide-band meeting data into the same space for improved model training. In this paper we focused on integration of discriminative adaptation into this scheme. This integration is not straightforward and we present the complexity of this process. The models are tested on the NIST RT'05 meeting evaluation where a relative reduction in word error rate of 5.6% against non-adapted meeting system was achieved.

Index Terms: Speech recognition, Discriminative training, LVCSR, Model adaptation, CMLLR

1. Introduction

The amount of training data has a crucial effect on the accuracy of HMM based meeting recognition systems but data in the meeting domain is still sparse and hence a common approach is to utilize other corpora for acoustic model training. One possibility to improve the system performance is to perform adaptation of models trained on considerably larger amounts of data. Typical domains with such large amounts of recorded material are broadcast news (BN) or conversational telephone speech (CTS). Depending on the domain difference one would try to adapt to either different recording environments or different speech type. CTS [1] is a good candidate for such adaptation – the typical speaking style matches that in meetings, but due to the telephone channel it is recorded with narrow bandwidth (NB, sampling frequency 8 kHz). Therefore an adaptation to meeting domain is not trivial, as the standard bandwidth for meeting recordings is 16 kHz (wide-band, WB).

The intuitive way to circumvent this problem is to downsample meeting data to NB and then adapt CTS models in that domain. This is however suboptimal as the upper band (4-8 kHz) was found to contain useful information[2]. The solution of this problem is to adapt CTS models to WB data using global Constrained Maximum Likelihood Linear Regression (CMLLR) [3] , which can be equivalently interpreted as feature space transformation performing a WB→NB conversion. With this approach, even though naturally the upper band information cannot be recovered for CTS data, we can still make use of the richer information in actual target domain recordings. This was already applied to meeting data in our previous work [4].

Once the WB features are rotated into the joint domain, it is possible to use any adaptation technique to adapt the CTS models into the transformed WB data. We used MAP adaptation [5]. In our implementation, it is applied iteratively, so output HMMs from previous iteration are taken as a prior for the current iteration [2]. This approach allows to give better state alignment and smoother convergence. There is however, a risk of overtraining, so the optimal adaptation control value has to be set higher than in standard MAP, and the number of iterations used also controls the adaptation process performance.

The basic idea of this process is showed in the figure 1. Upper branch shows traditional downsampling and lower branch presents our approach.

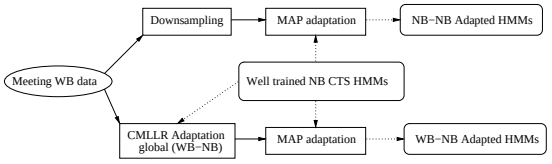

Figure 1: Downsampled and WB→NB adapted system.

An alternative to CMLLR is a multiple MLLR transforms governed by regression class trees. We experimented also with this type of adaptation: similar results can be obtained, however integration of MLLR with advanced techniques such as HLDA, SAT, . . . is infeasible because MLLR adaptation does not have a feature domain interpretation such as CMLLR.

In the following sections, implementation details on HLDA and SAT will be presented. Next, the paper will discuss training with focus on finding optimal CTS prior for discriminative adaptation.

2. WB→NB transform in HLDA estimation

Heteroscedastic Linear Discriminant Analysis (HLDA)[6], which is also in common use in speech recognition systems, provides a linear transformation that can de-correlate features

This work was partly supported by European project AMIDA (FP6-033812), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 201/2006. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383. Thanks to Cambridge University Engineering Department making the h5train03 CTS training set available.

and reduce the dimensionality while preserving the discriminative power of the features.

The computation of the HLDA matrix requires to collect full-covariance statistics assigned to particular class. In our work, the classes are given by Gaussian mixture components.

In the WB→NB adapted systems, the HLDA can be taken from CTS prior models. But taking only CTS data would limit discrimination power on meeting data. Thus it is important to make use of the meeting data for estimation of the HLDA matrix as well. Therefore full covariance estimates from both sets are collected and combined in a MAP-style adaptation of the statistics. The CTS full-covariance statistics ($\mathbf{\Sigma}^{(m)}_{(CTS)}, \boldsymbol{\mu}^{(m)}_{(CTS)}, \gamma^{(m)}_{(CTS)}$) are considered as priors and the WB→NB transformed WB statistics ($\hat{\mathbf{\Sigma}}^{(m)}_{(WB)}, \hat{\boldsymbol{\mu}}^{(m)}_{(WB)}, \hat{\gamma}^{(m)}_{(WB)}$) are taken for the adaptation. More details can be found in [4].

3. WB→NB transform in Speaker Adaptive Training

Speaker adaptive training (SAT) is a technique used to suppress cross-speaker variance [7]. The implementation in [3], used for this work, requires estimation of a set of CMLLR transforms to adapt speaker dependent training data to a global model. These transforms are then used during main model training.

Due to the feature space interpretation of CMLLR an implementation of WB→NB transforms incorporated in SAT training is a straightforward procedure. The model training is replaced by adaptation of so-called prior model with an application of a set of transforms. The whole procedure can be described as follows:

1. Choose CTS HLDA prior model.
2. Rotate the WB data by WB→NB CMLLR transform and project those into the MAP-HLDA space.
3. Use the prior to estimate SAT CMLLR transforms for each speaker in the training data $\hat{\mathbf{o}}(t)$.
4. Take the prior and run iterative MAP using the rotated data transformed by the respective SAT CMLLR transform.
5. Estimate a new set of SAT CMLLR transforms using the final models and go to step 4.

This process can be repeated iteratively until performance starts to degrade. In our experiments no improvement was noticed after the second iteration.

4. Discriminative training of WB→NB adapted system

The discriminative approaches are getting widely used in training of acoustic models for state-of-the-art recognition systems. We decided to improve our system by using discriminative MAP adaptation. Several discriminative criteria are available but usually best performance is achieved by using the Minimum Phone Error (MPE) criterion. MPE-MAP adaptation introduced in [8] is iterative process, where each iteration consists of two steps: First, a given prior model is adapted using standard (ML-)MAP adaptation. However, the resulting model is used only as a prior for the following MPE update, where the parameters of the current model are shifted to make compromise between improving MPE objective function and obeying the prior distribution. Therefore, we need to distinguish two models that serve as the input for MPE-MAP adaptation: the (fixed) prior model and the starting point model, which is to be iteratively updated. It is usual practice to set the starting point to be equal to the prior. However, the problem for the practical implementation of WB→NB system lies in quite significant difference between the CTS prior models and WB→NB rotated adaptation data. Therefore, we first adapt CTS prior to rotated adaptation data using iterative ML-MAP [1] to obtain good starting point, which is further iteratively adapted using MPE-MAP (still with CTS model fixed as the prior). Although, each MPE-MAP iteration also contains single iteration of ML-MAP adaptation, performing the iterative ML-MAP prior to starting the discriminative adaptation turned out to be essential for successful use of MPE-MAP.

Data set	Size [h]
ctstrain03	278
ctstrain07sub	1000
ctstrain07	2000

Data set	Size [h]
ihmtrain05	112
ihmtrain07	183

Table 1: Used corpora and amounts of data.

5. Experimental Setup

5.1. Data

CTS models were trained on two data sets. The AMIDA ctstrain03 set is based on the h5train03 training set defined at Cambridge University. It consists of Switchboard1, Switchboard2 and Call Home English data. Sentences containing words, which do not occur in the training dictionary were removed. The total amount of CTS training data was 278 hours.

The ctstrain03 set was further extended by data from the Fisher 1 and 2 corpora. The resulting ctstrain07 data set was comprised of 2000 hours of data. Previous work on training system from large databases [9] showed no yield over 1000 hours using ML training but the discriminative training techniques were still improving system performance significantly. Therefore, a smaller ctstrain07sub set containing 1000h was also defined with the requirements of complete speaker coverage and similar distribution to the full set.

Meeting training data has also consisted of two parts. The ihmtrain05 set (IHM stands for independent head-set microphone) was defined for AMI RT05 Rich Transcription system [10]. It contained 112h of close talk speech from ICSI (73 hours), NIST (13 hours), ISL (10 hours) and AMI (16 hours) corpora.

With release of the full AMI corpus[2], the amount of available meeting data increased. Extension of the ihmtrain05 set with the AMI corpus and new recordings from NIST resulted in the ihmtrain07 set with a total amount of 183 hours.

2 hours of NIST RT05 IHM data were taken for testing. The speech/non-speech segmentation was taken from NIST references and results were obtained by acoustic rescoring of lattices from AMI NIST 2005 Rich Transcription system [10]. All results will be presented as word error rates (WER).

5.2. System description

The speech recognition system is based on HMM cross-word tied-states triphones. MF-PLP features were generated using the HTK implementation, with a total number of 13 coefficients. Deltas, double-deltas and in the HLDA system, also triple-deltas were added, so that the feature vector had 39 and 52 dimensions respectively. Cepstral mean and variance nor-

[1] In contrary to iterative MAP described in section 1 the prior does not change over the iterations and stay fixed to CTS model

[2] Information on AMI corpus is available at `http://corpus.amiproject.org`

	WB→NB CMLLR	
	CTS03→ihm05	CTS07sub→ihm07
CTS03 52d	36.3	-
CTS07sub 52d	35.4	34.0

Table 2: CTS 52d models: Effect of WB→NB CMLLR and training data size. Tested by acoustic rescoring of rt05 lattices.

Train set	ctstrain03	ctstrain07sub	ctstrain07
CTS SAT ML	31.3	29.6	29.6
CTS SAT MPE	28.0	26.4	25.9

Table 3: CTS system: Dependency of WER on training data size obtained by acoustic rescoring of eval01 lattices. Systems were adapted to the test speakers in all cases.

malization was applied with the mean and variance vectors estimated on each meeting channel. HLDA was estimated with Gaussian components as classes and the dimensionality was reduced to 39. VTLN warping factors were applied by adjusting the centres of the Mel-filterbanks.

6. CTS system development

Due to smaller set size the main development was investigated with adaptation of ctstrain03 models to ihmtrain05 data. When the optimal configuration was found, ctstrain07 models were adapted on ihmtrain07 data.

Ctstrain03 models were trained from scratch using mixture-up training. The final models contained ≈7600 tied states and 16 Gaussian mixtures per state. Ctstrain07sub models were bootstrapped from ctstrain03 models, decision tree clustering produced ≈10000 tied-states and mixture-up training produced 20 Gaussian components per state.

These models were retrained using single pass retraining in 52 dimensional space and WB→NB$_{cts07sub\rightarrow ihm07}$ global CMLLR transform was estimated. Table 2 shows improvement given by more data for CTS training and WB→NB transform. The results were generated by direct decoding of meeting data using just CTS models and WB→NB transform, therefore no parameters were re-estimated.

The HLDA transform matrix was estimated with respect to further adaptation to the meeting domain. Full covariance statistics were collected for both data sets and merged using MAP criteria (see section 2). The model parameters were projected into the new space and further trained using SAT. Next, the final SAT models were trained discriminatively using the MPE criterion on the full 2000h training set.

Table 3 shows the effect of amount of training data. As expected ML training on 1000 hours does not give any improvement over that on 2000h (note that the decision trees were not re-done for the larger set though). MPE training using 2000h shows a substantial gain of 3.7% absolutely against ML and 0.5% compared to 1000h MPE models. All model sets makes use MAP-HLDA. As these are SAT models speaker based adaptation was applied in all test cases.

7. Adapting CTS models to meeting data

In this section, the experiments are described where CTS models are adapted to WB→NB rotated meeting data. As implementation follows the description in section 3, all the models described in this section make use of MAP-HLDA and SAT.

System	Adaptation	WER
IHM05 WB SAT	CMLLR$_{SAT}$	27.5
IHM05 NB SAT	CMLLR$_{SAT}$	28.8
CTS03 NB-NB SAT	CMLLR, CMLLR$_{SAT}$	27.9
CTS03 WB-NB SAT	CMLLR$_{WB\rightarrow NB}$, CMLLR$_{SAT}$	26.5

Table 4: Results of HLDA SAT systems.

Prior	Starting point	Adaptation	WER [%]
CTS03_MPE	CTS03_MPE	*MPE-MAP*	27.2
CTS03_MPE	-	*ML-MAP*	27.0
CTS03_MPE	CTS_ML-MAP	*MPE-MAP*	25.6

Table 5: MPE-MAP: Effect of selecting different prior and starting point models.

7.1. Adaptation in the NB domain by down-sampling

To see a comparison of the proposed approach and traditional downsampling, meeting data was downsampled and CTS models were adapted into this domain using equivalent schemes to those WB→NB system training. We will refer to this system as to CTS03 NB→NB.

Table 4 presents the various SAT systems. The first two lines show results obtained with systems trained only on WB and downsampled NB meeting data (no CTS prior). A 1.3% loss of accuracy is caused by downsampling the data. The best performance, 26.5% WER absolute, was generated by the adapted CTS03 WB→NB SAT system which is a 3.3% relative improvement over the non-adapted IHM05 WB SAT system and 4.6% relative improvement over CTS03 NB-NB adapted system.

8. Discriminative training of WB→NB adapted system

8.1. Discriminative adaptation of WB→NB HLDA system

For simplicity, a first experiment with the discriminatively adapted models will start with do not make use of SAT, just MAP-HLDA. It is important for MPE-MAP adaptation to have a relevant prior information about the target domain distributions. Consequently, the CTS MAP-HLDA models described in previous sections were further trained using MPE to get a better prior model. In section 4, we also mentioned the importance of having proper model that serves as a starting point for MAP-MPE. Therefore, experiments were conducted exploring the influence of the starting point and the adaptation approach.

Firstly, we compared ML-MAP and MPE-MAP using CTS_MPE prior and different starting point models.

In Table 5 we can see that MPE-MAP using the CTS_MPE starting point does not give any improvement, but instead even 0.2% degradation of accuracy. This is most likely due to starting point models being too far from the target data. Consequently, we decided to use iteratively ML-MAP adapted models as the starting point for MPE-MAP adaptation. This approach yielded 1.6% absolute improvement compared to the results with using the CTS_MPE models .

8.2. Discriminative adaptation of the WB→NB HLDA SAT system

For these experiments the CTS03_SAT_MPE model from section 6 was used as prior for discriminative adaptation. When processing the meeting data, SAT transforms were estimated

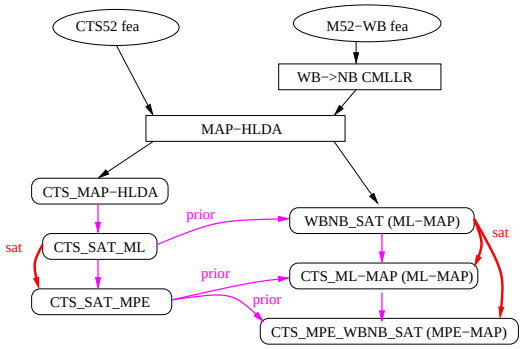

Figure 2: Adaptation scheme of MPE-MAP adaptation into the WB→NB features.

Prior	Starting point	Adaptation	WER[%]
CTS03_SAT_MPE	-	*ML-MAP*	25.7
-	CTS_ML-MAP	*MPE*	24.2
CTS_ML-MAP	CTS_ML-MAP	*MPE-MAP*	24.1
CTS03_SAT_MPE	CTS_ML-MAP	*MPE-MAP*	23.9

Table 6: MPE-MAP in the SAT: Effect of selecting the prior and starting point models.

based on the CTS03 WB-NB resulting models from section 7.1. The transforms remained fixed for further processing.

Using an equivalent setup as that in section 8.1, the CTS03_SAT_MPE models were adapted in the WB→NB rotated domain using iterative ML-MAP with application of the above SAT transforms. These models, further shortly referred to as CTS_ML-MAP, are used as the starting point for the final MPE-MAP adaptation (see the scheme in figure 2).

To investigate the effect of the CTS prior, the CTS_MPE_WB→NB_SAT_ML-MAP models were further trained using just MPE. Table 6 shows that a 1.5% absolute improvement is obtained by MPE training of ML-MAP adapted models. Incorporation of the CTS prior gives an additional 0.3% improvement. When using the CTS_ML-MAP models as the prior and starting point no significant gain was observed.

The final models were successfully used in the AMI LVCSR system for NIST 2007 Rich Transcription evaluation.

9. Final WB→NB adapted system

All experiments to adapt a CTS07 MPE SAT models into the WB→NB rotated domain used the same algorithm as described above.

First, an unadapted baseline system was trained just on the new meeting data (ihmtrain07) which yielded 1.7% absolute improvement in ML training over the ihmtrain05 system and more than 1% when using MPE (see Table 7).

To capitalize on these gains the CTS07 MPE SAT models were adapted in the WB→NB rotated domain according to the scheme in section 8.2. Therefore, first, MPE starting point models were trained using ML-MAP and MPE-MAP adaptation followed.

Table 8 shows a 1.8% absolute gain due to adding training data and 1.3% improvement by adaptation from CTS.

Data	ihmtrain05	ihmtrain07
ML SAT	27.5	25.8
MPE SAT	24.5	23.4

Table 7: Unadapted meeting systems: Dependency of WER on the train data size.

Adaptation	CTS03→ihmtrain05	CTS07→ihmtrain07
CTS SAT prior		
ML-MAP	26.5	25.1
CTS SAT MPE prior		
ML-MAP	25.7	23.8
MPE-MAP	23.9	22.1

Table 8: WB → NB: Effect of training data and adaptation approach.

10. Conclusion

We successfully implemented an adaptation technique where WB data is transformed to the NB domain by CMLLR feature transform. Here, the well trained CTS models are taken as prior for adaptation. A solution on how to apply this transform for HLDA and SAT systems was given using maximum likelihood where a 4.6% relative improvement against adaptation in the downsampled domain was obtained. Next, ML-MAP was replaced by the discriminative MPE-MAP scheme, where a 2.4% relative improvement over the non-adapted meeting system was shown.

The Fisher corpora were included for improving of the CTS prior model and also some new meeting data resources. In the final MPE-MAP implementation, we obtained a 5.6% relative improvement over non-adapted meeting system.

11. References

[1] A. S. et al., "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaulation system," in *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop*, Edinburgh, UK, July 2005.

[2] T. Hain, J. Dines, G. Gaurau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of conference room meetings: an investigation," in *In Proceedings of Interspeech 2005*, Lisabon, Portugal, 2005.

[3] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," 1997. [Online]. Available: citeseer.ist.psu.edu/gales97maximum.html

[4] M. Karafiat, L. Burget, T. Hain, and J. Cernocky, "Application of cmllr in narrow band wide band adapted systems," in *Proc. INTERSPEECH 2007*. International Speech Communication Association, 2007, p. 4.

[5] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[6] M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[7] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP '96*, vol. 2, Philadelphia, PA, 1996, pp. 1137–1140. [Online]. Available: citeseer.ist.psu.edu/anastasakos96compact.html

[8] D. Povey, M. Gales, D. Kim, and P. Woodland, "Mmi-map and mpe-map for acoustic model adaptation," in *Proc. Eurospeech 2003*, no. ISSN 1018-4074, Geneva, Switzerland, 2003.

[9] G. Evermann, H. Chan, M. Gales, B. Jia, D. Mrva, P. Woodland, and K. Yu, "Training lvcsr systems on thousands of hours of data," in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal processing (ICASSP)*, no. ISSN: 1520-6149, Philadelphia, PA, USA, march 2005, pp. Volume I, pp. 209–212.

[10] T. H. et al., "The 2005 AMI system for the transcription of speech in meetings," in *Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop*, Edinburgh, UK, July 2005.