Application of CMLLR in narrow band wide band adapted systems

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

[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. Conversational telephone speech matches speech in meetings well. However it is naturally recorded with low bandwidth. In this paper we present a scheme that allows to transform wide-band meeting data into the same space for improved model training. The transformation into a joint space allows simpler and more efficient implementation of joint speaker adaptive training (SAT) as well as adaptation of statistics for heteroscedastic discriminant analysis (HLDA). Models are tested on the NIST RT'05 meeting evaluation where a relative reduction in word error rate of 4% was achieved. With the use of HLDA and SAT the improvement was retained.

Index Terms: Speech recognition, Speech processing, LVCSR, Speech adaptation, CMLLR

1. Introduction

The amount of training data is critical in speech-to-text transcription systems based on the Hidden Markov Models (HMM). Our working area is speech recognition of meeting data. 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 large amounts of recorded material are broadcast news (BN) or conversational telephone speech (CTS). Depending on the domain difference once would try to adapt to either different recording environments or different speech type. As the type of speech is often the cause for greater variability adaptation of similar speech types is generally preferred. Hence for the meeting domain adaptation of models trained on CTS data is appropriate [1].

This however is not necessarily trivial as data recorded over telephone is necessarily band-limited to 8kHz (narrow-band,NB), whereas the standard bandwidth for meeting recordings is 16kHz (wide-band,WB).

The standard way to cope with this problem is to down-sample meeting data to NB and to adapt CTS models in that domain. In previous work we found that this process is suboptimal [2]. There it was show that a global feature space transformation based on Maximum Likelihood Linear Regression (MLLR) [3] to perform a NB to WB conversion and applied it on CTS models can yield improved performance of both down-sampling and unadapted training [2]. An iterative Maximum a Posteriori (MAP) adaptation [4] is followed to settle the adapted CTS models to meeting data.

But MLLR is difficult to implement with advanced techniques such as Heteroscedastic Linear Discriminative Analysis (HLDA) [5], Speaker Adaptive Training (SAT) [12].

In this paper we propose a single constrained MLLR [6] for feature space transformation from WB to NB. This method is straight-forward in implementation and requires little extra computation in decoding. However, state-of-the-art systems in meeting transcription [7] are using other techniques such as heteroscedastic discriminant analysis (HLDA), speaker adaptive training (SAT), or discriminative model parameter estimation to derive model sets. In all of these techniques accurate and consistent statistics derived from both adaptation and baseline training set are required. In the case of bandwidth difference the merge of statistics has to include adaptation to the feature domain.

2. Transformation between wide-band and narrow-band

As introduced above the CMLLR transformation is estimated to adapt CTS models to meeting WB data. This can equally be interpreted as a projection of WB meeting data into NB CTS domain. This does not seem an obvious choice as one constrains the increased richness of WB meeting data. However, the alternative, i.e. transforming NB CTS data into the WB space, clearly can only add distortion, but add no information. Hence better model training on the larger amounts of data is given priority. Using a transformation matrix to make meeting data more like CTS data may however preserve some of the characteristics only visible with higher bandwidth.

Then, the WB to NB transformation is applied on the WB train features. Now, it is possible to use any adaptation technique to adapt the CTS models into the transformed space.

In our experiments, we use MAP adaptation applied iteratively, so that output HMMs from previous iteration are taken as prior for current iteration [2]. The first iteration is classical MAP. The CTS prior is used to align transformed WB data, which is not very accurate. For further iterations, the MAP prior at current iteration is taken from previous iteration. It gives better full Gaussian alignment and smooth convergence to transformed WB system. There is however a risk of overtraining, so the optimal τ value should be set higher than during common MAP, the number of iterations then provides a better adaptation control.

3. Heteroscedastic Linear Discriminant Analysis

Linear discriminant analysis is a standard technique that allows decorrelation of the feature space while rotating to achieve maximum discrimination. However, the basic assumption is that classes have equal covariances matrices. Heteroscedastic LDA alleviates this constraint [9]). Both schemes allow to reduce dimensionality. For HLDA, each feature vector used to derive the transformation must be assigned to a class. When performing the dimensionality reduction, HLDA allows to preserve such dimensions, where feature vectors representing individual classes are best separated. An efficient iterative algorithm [10, 11] is used in our experiments to estimate HLDA matrix. 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.

3.1. WB→NB transform in HLDA estimation

The easiest way to train WB→NB HLDA system is to take HMMs and HLDA matrix already trained on CTS data and use them as prior for adaptation of the WB→NB transformed features, the same way as in section 2.

It is important to estimate statistics from both data sets, to take an advantage of the meeting data for HLDA matrix estimation. We use MAP adaptation of statistics, so the CTS full-covariance statistics $(\boldsymbol{\Sigma}^{(m)}_{(CTS)}, \boldsymbol{\mu}^{(m)}_{(CTS)}, \gamma^{(m)}_{(CTS)})$ are considered as priors and the WB→NB transformed WB statistics $(\hat{\boldsymbol{\Sigma}}^{(m)}_{(WB)}, \hat{\boldsymbol{\mu}}^{(m)}_{(WB)}, \hat{\gamma}^{(m)}_{(WB)})$ are taken for the adaptation. The transformation of WB statistics is given by:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^{(m)}_{(WB)} &= \mathbf{A}_{(WB\rightarrow NB)}\boldsymbol{\mu}^{(m)}_{(WB)} + \mathbf{b}_{(WB\rightarrow NB)}, && (1)\\
\hat{\boldsymbol{\Sigma}}^{(m)}_{(WB)} &= \mathbf{A}_{(WB\rightarrow NB)}\boldsymbol{\Sigma}^{(m)}_{(WB)}\mathbf{A}_{(WB\rightarrow NB)}{}^{T} && (2)\\
\hat{\gamma}^{(m)}_{(WB)} &= \gamma^{(m)}_{(WB)}, && (3)
\end{aligned}$$

where $\mathbf{A}_{(WB\rightarrow NB)}$ and $\mathbf{b}_{(WB\rightarrow NB)}$ are given by $WB{\rightarrow}NB$ CMLLR transform.

The estimation of an arbitrary covariance matrix $\boldsymbol{\Sigma}^{(m)}$ is given by:

$$\boldsymbol{\Sigma}^{(m)} = \frac{\sum_{t=1}^{T}\gamma^{(m)}(t)\mathbf{o}(t)\mathbf{o}(t)^{T}}{\gamma^{(m)}} - \boldsymbol{\mu}^{(m)}\boldsymbol{\mu}^{(m)^{T}}, \quad (4)$$

where $\gamma_{(m)}$ is global occupation count of component m given by:

$$\gamma^{(m)} = \sum_{t=1}^{T}\gamma^{(m)}(t) \quad (5)$$

According to equation 4, the matrix $\boldsymbol{\mu}^{(m)}\boldsymbol{\mu}^{(m)^{T}}$ has to be added to "de-normalize" the covariance matrix, and get back the original "scatter" matrix. Here, the merging (adaptation) of covariance matrices is possible. Therefore, the MAP adaptation of the statistics is given by:

$$\begin{aligned}
\breve{\boldsymbol{\mu}}^{(m)} &= \frac{\hat{\gamma}^{(m)}_{(WB)}\hat{\boldsymbol{\mu}}^{(m)}_{(WB)} + \tau\boldsymbol{\mu}^{(m)}_{(CTS)}}{\hat{\gamma}^{(m)}_{(WB)} + \tau} && (6)\\
\breve{\boldsymbol{\Sigma}}^{(m)} &= \frac{(\hat{\boldsymbol{\Sigma}}^{(m)}_{(WB)} + \hat{\boldsymbol{\mu}}^{(m)}_{(WB)}\hat{\boldsymbol{\mu}}^{(m)^{T}}_{(WB)})\hat{\gamma}^{(m)}_{(WB)}}{\hat{\gamma}^{(m)}_{(WB)} + \tau} + && (7)\\
&\quad + \frac{(\boldsymbol{\Sigma}^{(m)}_{(CTS)} + \boldsymbol{\mu}^{(m)}_{(CTS)}\boldsymbol{\mu}^{(m)}_{(CTS)}{}^{T})\tau}{\hat{\gamma}^{(m)}_{(WB)} + \tau} - \breve{\boldsymbol{\mu}}^{(m)}\breve{\boldsymbol{\mu}}^{(m)^{T}}\\
\breve{\gamma}^{(m)} &= \gamma^{(m)}_{(CTS)} && (8)
\end{aligned}$$

where τ is the a control constant and $\breve{\boldsymbol{\mu}}^{(m)}, \breve{\boldsymbol{\Sigma}}^{(m)}, \breve{\gamma}^{(m)}$ are the resulting adapted statistics.

In the next step, HLDA is estimated from these statistics and HMMs are updated by projecting the statistics through HLDA. The standard iterative MAP adaptation follows to settle updated HMMs in the same way as in section 2.

4. WB→NB transform in Speaker Adaptive Training

Speaker adaptive training (SAT) is a technique used to suppress cross-speaker variance [12]. The implementation in [6], which we used, estimates a set of CMLLR transforms to adapt speaker dependent training data to a global model. These transforms are used during the training.

An implementation of WB→NB transform during the SAT training is a straightforward procedure:

1. The WB data are rotated into the NB space.
2. The SAT trained CTS HMMs are used to estimate a set of SAT CMLLR transforms on WB→NB rotated data.
3. The SAT CTS HMMs are further adapted using iterative MAP to WB→NB rotated data with apply SAT CMLLR transforms (So, the WB data are rotated twice, first by global WB→NB transform,secondly by speaker dependent SAT transform).

5. EXPERIMENTS

The data used for train CTS models are based on h5train03 training set defined at Cambridge University. Sentences containing words, which do not occur in training dictionary were removed. The total amount of CTS training data is 270 hours.

Meeting WB data contains 112h of close talk speech from ICSI (73 hours), NIST (13 hours), ISL (10 hours) and AMI (16 hours) corpora.

The speech recognition system is based on HMM cross-word tied-states triphones. Mel-PLP's features were generated in classical way, the resulting number of coefficients is 13. Deltas, double- and in HLDA system also triple-deltas were added, so that the feature vector had 39 respectively 52 dimensions. Cepstral mean and variance normalization was applied with the means and variances estimated on each meeting channel. HLDA was estimated with Gaussian components as classes to reduce the dimensionality to 39. VTLN warping factors were applied on filters of Mel filterbanks.

To get better comparison three baselines were trained:

- Non-adapted WB systems (WB basic, WB HLDA, WB HLDA SAT) trained on the WB meeting data only.

- Non-adapted NB systems (NB basic, NB HLDA, NB HLDA SAT) trained on the downsampled meeting data only.
- Adapted NB-NB systems (NB-NB basic, NB-NB HLDA, NB-NB HLDA SAT) based on adaptation of CTS models to downsampled meeting train data using same techniques as WB→NB systems.

All results were obtained by acoustic rescoring of lattices from AMI NIST 2005 Rich Transcription system on NIST RT05 ihm evaluation data [13]. The segmentation was taken from NIST references.

5.1. WB→NB Basic system

The initial CMLLR WB→NB transform was estimated using CTS models and WB data. This is, however, not too accurate due to data mismatch. Therefore, the process was run iteratively and next iterations use CMLLR WB→NB from previous iteration to get better full state alignment. We use one block diagonal transform. We also tried to run experiments with more complex transform structure (speech/silence). They gave similar results but with significantly increased complexity.

Table 1 shows performance of CTS models directly applied on WB→NB rotated test data. After 16 CMLLR iterations we do not have any improvement therefore this transformation was fixed for the following experiments.

CMLLR iteration	WER [%]
4	40.0
8	35.7
16	35.4
20	35.4

Table 1: WB→NB - Performance of CTS models on WB meeting with different WB→NB CMLLR quality.

Next, the WB data are projected into NB space and iterative MAP adaptation follows in same way as was described in section 2.

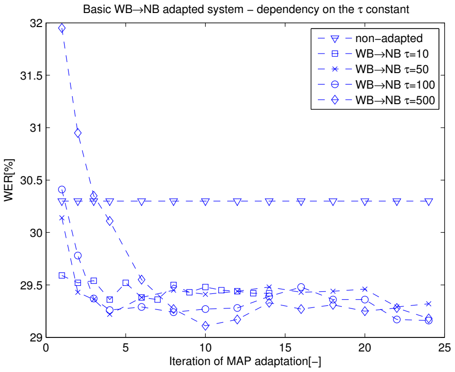

Figure 1: WB→NB - τ constant in the iterative MAP with fixed number of 16 CMLLR iterations.

Figure 1 shows the performance with different τ values and fixed number of 16 CMLLR iterations. We see that value τ is not too important, but higher values give more smooth convergence to the best WER, after it the system tends to be overtrained. The τ was fixed on value 50 for further experiment.

Train set	Adaptation	WER [%]
WB	none	30.3
NB	none	30.7
NB-NB	CMLLR MAP	29.8
WB-NB	CMLLR$_{WB\rightarrow NB}$,MAP	29.0

Table 2: Performance of basic systems

The best performance of basic WB→NB system is 29% which is a 4.4% relative improvement over the non-adapted WB system and 2.7% over NB-NB adapted system (see Table 2).

5.2. WB→NB HLDA experiments

HLDA transform is used to perform dimensionality reduction from 52 dimensional space to 39. Therefore, WB→NB CMLLR has to be also 52 dimensional. To be able to estimate this transform, 52 dimensional CTS models were trained.

Full covariance statistics have to be accumulated for both data sets to estimate HLDA matrix. For the merging procedure, it is important to collect them by the same clustered models. The WB→NB transform was applied on WB data and all statistics were collected with the CTS models. Consequently, the WB statistics were collected in rotated space, thus equations 1,2 did not need to be applied.

Equations 6,7,8 were used to merge the statistics according to MAP criterion. The HLDA matrix was estimated and CTS HMMs were updated. Further, the iterative MAP was applied to settle the HMMs into the new space.

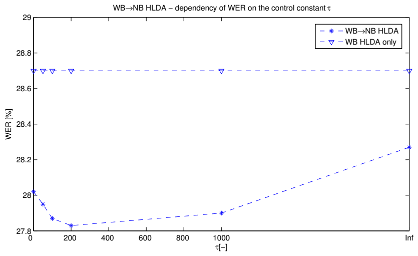

Figure 2: WB→NB HLDA - the τ value in a MAP adaptation of statistics.

Figure 2 shows the dependency of WER on the τ value during MAP merging of statistics. The system generates the best accuracy 27.8% with τ = 200. It gives a 3% relative improvement over the non-adapted HLDA system (see Table 3).

The τ = "Inf" means that no WB data was used to estimate HLDA. This is an interesting result because it still gives 1.5% relative improvement even if HLDA has not been trained on WB data at all. It means that the accuracy of the prior HMM has higher importance than the kind of data used for HLDA estimation.

Table 3 shows two kinds of HLDA estimation in NB-NB adapted system. First, the HLDA were taken directly from CTS system. Secondly, the HLDA were computed over the MAP adapted statistics like in section 3.1. The CTS statistics were

System	HLDA adaptation	WER [%]
WB	none	28.7
NB	none	29.7
NB-NB	HLDA taken from CTS	28.6
NB-NB	MAP HLDA	29.0
WB-NB	CMLLR$_{WB \rightarrow NB}$, MAP HLDA	27.8

Table 3: Performance of HLDA systems

System	Adaptation	WER [%]
WB	none	27.5
NB	none	28.8
NB-NB	CMLLR, CMLLR$_{SAT}$	27.9
WB-NB	CMLLR$_{WB \rightarrow NB}$, CMLLR$_{SAT}$	26.6

Table 4: Results of HLDA SAT systems

taken as a prior and statistics collected over the downsampled meeting data were used for adaptation. It is interesting that HLDA taken from CTS for NB-NB adapted system gives better performance than adapted HLDA. It seems that statistics collected over the downsampled meeting data does not contain any additional information for HLDA estimation.

5.3. WB→NB HLDA SAT experiments

Speaker adaptive training was used in addition to HLDA system. The implementation is similar to section 4 but extended for HLDA implementation:

1. Take the CTS HLDA prior model
2. Rotate the WB data by:

$$\hat{\mathbf{o}}(t) = \mathbf{A}_{(HLDA)}\mathbf{A}_{(WB \rightarrow NB)}\mathbf{o}(t) + \mathbf{A}_{(HLDA)}\mathbf{b}_{(WB \rightarrow NB)}, \tag{9}$$

   where $\mathbf{A}_{(HLDA)}$ is HLDA matrix.
3. Use the prior to estimate an independent SAT CMLLR transform for each speaker in the training data $\hat{\mathbf{o}}(t)$.
4. Take the prior and run iterative MAP with the rotated data $\hat{\mathbf{o}}(t)$ transformed by the respective SAT CMLLR transform.

As is shown at Table 4, the best performance of WB→NB in HLDA SAT system is 26.6% which is a 3.3% relative improvement over the non-adapted HLDA SAT system and 4.6% relative improvement over NB-NB adapted system.

6. Conclusion

We successfully implemented an adaptation technique where WB data are rotated into the NB domain by CMLLR feature transform. Here, the CTS well trained models are taken as prior for adaptation. A solution to apply this transform in HLDA and SAT systems was given. On NIST RT 2005 task, the relative improvement was higher than 4% in the basic system and more than 3% in HLDA and SAT systems.

Our current work in acoustic modeling for LVCSR concentrates on the use of WB→NB transform along with discriminative and minimum phoneme error (MPE) training and adaptation.

7. Acknowledgments

partly supported by European project AMI (Augmented Multiparty Interaction, FP6-506811) and Grant Agency of Czech Republic under project No. 102/05/0278. The hardware used in this work was partially provided by CESNET under project No. 119/2004, No. 162/2005 and 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.

8. References

[1] Andreas Stolcke 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.J.F. Ordelman, and S. Renals, “Transcription of conference room meetings: an investigation,” in *In Proceedings of Interspeech 2005*, Lisabon, Portugal, 2005.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” .

[4] 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.

[5] N. Kumar and A. G. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition,” *Speech Communication*, pp. 283–297, 1998.

[6] M. Gales, “Maximum likelihood linear transformations for hmm-based speech recognition,” 1997.

[7] T. Hain et al., “The AMI system for the transcription of speech meetings,” in *In Proceedings of ICASSP 2007*, Honolulu, Hawai USA, April 2007.

[8] V. Digalakis and L. Neumeyer, “Speaker adaptation using combined transformation and bayesian methods,” in *Proc. ICASSP ’95*, Detroit, MI, 1995, pp. 680–683.

[9] N. Kumar, *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, Ph.D. thesis, John Hopkins University, Baltimore, 1997.

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

[11] L. Burget, “Combination of speech features using smoothed heteroscedastic linear discriminant analysis,” in *8th International Conference on Spoken Language Processing*, Jeju island, KR, oct 2004.

[12] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, “A compact model for speaker-adaptive training,” in *Proc. ICSLP ’96*, Philadelphia, PA, 1996, vol. 2, pp. 1137–1140.

[13] T. Hain 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.