Vysoké učení technické v Brně

Fakulta informačních technologií

Ústav počítačové grafiky a multimédií

Ing. Martin Karafiát

Aplikace lineárních transformací pro trénování systémů rozpoznávání spojité řeči s velkým slovníkem adaptovaný napříč doménami

Study of Linear Transformations Applied to Training of Cross-Domain Adapted Large Vocabulary Continuous Speech Recognition Systems

Obor: Informační technologie

Zkrácená verze disertační práce

Školitel: Jan Černocký

Oponenti:

Datum obhajoby:

Klíčová slova: LVCSR systém, meeting recognition, linearni transformace, Adaptace, HLDA, CMLLR, MLLR.

Keywords: LVCSR system, meeting recognition, linear transform, Adaptation, HLDA, CMLLR, MLLR, narrow band - wide band.

Rukopis disertační práce je uložen na Fakultě informačních technologií Vysokého učení technického v Brně, Božetěchova 2, 61266 Brno. Plný text disertační práce je k disposici na:
`http://www.fit.vutbr.cz/˜karafiat/publi/dis.pdf`

ISBN 80-214-

ISSN 1213-4198

Contents

Chapter 1

Introduction

In last few years, significant importance was put on audio/visual data. Two different kinds of information are combined here. The video records (seeing modality) is useful source of information but it is quite difficult to find interesting points. The audio records (hearing modality) can be easily processed by speech recognition system. The recognition output can be further used by indexer or as an input for other information retrieval techniques: summarization, spoken term detection…

Due to two modalities on the input, this approach is also called "multimodal approach" - a weakness of one modality could be complemented by strength of other. Obviously, the complexity of such system can be huge and it can touch many science branches. In this work, we focus only on one part of the whole system - speech recognition system.

The used data were taken from "meeting sessions". We can imagine that a few more or less intelligent people are sitting around the table and spontaneously discuss some technical problem. Obviously, the language has to be same but dialects can differ and the speakers are frequently non-native. The whole discussion is recorded by video camera and microphones.

1.1 Meeting recognition

In comparison to telephone conversations, the meetings speech differs in channel. Different microphones are used and the bandwidth is different. Telephone speech is naturally recorded in low bandwidth (8kHz) and meetings are recorded in wider band (16kHz). But the meeting speech is quite similar to CTS in conversation style, therefore similar problems exists. A big proportion of non-native speakers and different channel parameters increase the demands for relevant acoustic data[1]. Recordings of this kind of conversation took place at several sites (M4/AMI/AMIDA series of projects[2]. . .) but still not enough data is available.

1.2 Goals of this thesis

The aim is to increase the robustness of acoustic modeling techniques in meeting speech recognition, especially by the use of Heteroscedastic Linear Discriminant Analysis (HLDA). We propose MAP-Smoothed and Silence Reduced HLDA modifications.

Further, we focused on effective porting of telephone speech data resources into the meeting domain. A common problem is different bandwidth. The standard approach is to downsample the wide-band data, which is not too efficient due to loss of information from the upper band. We investigate substitution of the downsampling by adaptation which does not remove any information. Next, we focused on using this approach together with advanced techniques like HLDA, Speaker Adaptive Training (SAT) and discriminative training. The solution is not trivial, so mathematical development and extensive experimental work is presented.

[1] The data required for the training of language model can be partly derived from in-domain texts.

[2] www.m4project.org, www.amiproject.org

Chapter 2

Linear transforms in feature-space

In our experiments, linear transforms are used to decorrelate and reduce dimensionality of features.

2.1 Introduction into Heteroscedastic Linear Discriminant Analysis

The Heteroscedastic Linear Discriminant Analysis (HLDA) [6] can be used to derive linear projection de-correlating feature vectors and performing the dimensionality reduction. For HLDA, each feature vector that is used to derive the transformation must be assigned to a class. When performing the dimensionality reduction, HLDA allows to preserve useful dimensions, in which feature vectors representing individual classes are best separated. HLDA allows to derive such projection that best de-correlates features associated with each particular class [6, 3].

To perform de-correlation and dimensionality reduction, n-dimensional feature vectors are projected into first $p < n$ rows, $\mathbf{a}_{k=1...p}$, of $n \times n$ HLDA transformation matrix, $\mathbf{A}$. An efficient iterative algorithm [3] is used in our experiments to estimate matrix $\mathbf{A}$.

In our experiments, the classes are defined by each Gaussian mixture component m of each state s. The selection, that feature vector $\mathbf{o}(t)$ belongs to class j, is given by the value of occupation probability $\gamma_j(t)$.

2.2 Study of the HLDA

HLDA estimation algorithm requires estimation of full covariance statistics for each class. A Gaussian is usually considered as the class in a standard HMM system. This can however lead to noisy estimation of statistics even in case of well tuned system. Therefore, a smoothing techniques will be introduced in this chapter to obtain more robust HLDA estimation.

In our experiments, we added the third derivatives into the PLP feature stream, which gave us 52 dimensional feature vectors. **HLDA** transform was then trained to perform the projection from 52 to 39 dimension. The statistics were projected into the new space and HMM models were updated. A few additional Baum-Welch iterations were run to better settle HMM into the new space.

In chapter 3, we will investigate using of CTS models in meeting system, therefore the development run on both tasks. It is important to know if the best approach for telephone speech generalizes also for meetings. The testing was performed on eval01 test set (for CTS system) and on rt05 test set (for meeting system). For meeting system, VTLN was applied in advance.

2.2.1 Smoothed HLDA - SHLDA

SHLDA is a technique based on combination HLDA and LDA proposed in [1], where class covariance matrices are estimated more robustly, and at the same time, (at least the major) differences between covariance matrices of different classes are preserved. Smoothed HLDA (SHLDA) differs from HLDA only in the way of estimating of class covariance matrices. In the case of SHLDA, the estimate of class covariance matrices is given by:

$$\check{\boldsymbol{\Sigma}}_j = \alpha \hat{\boldsymbol{\Sigma}}_j + (1-\alpha)\boldsymbol{\Sigma}_{WC} \qquad (2.1)$$

where $\check{\boldsymbol{\Sigma}}_j$ is “smoothed” estimate of covariance matrix of class j. $\hat{\boldsymbol{\Sigma}}_j$ is original estimate of covariance matrix, $\boldsymbol{\Sigma}_{WC}$ is estimate of within-class covariance matrix and α is smoothing factor — a value in the range of 0 to 1. Note that for α equal to 0, SHLDA becomes LDA and for α equal to 1, SHLDA becomes HLDA.

2.2.2 MAP smoothed HLDA - MAP-SHLDA

SHLDA gives more robust estimation than standard HLDA but optimal smoothing factor α depends on the amount of data for each class. In extreme case, α should be set to 0 (HLDA) if infinite amount of training data is available. With decreasing amount of data, optimal α value will slide up to LDA direction.

To add more robustness into the smoothing procedure, we defined maximum a posteriori (MAP) smoothing similar to classical MAP adaptation of Gaussian parameters introduced in [4]. The within-class covariance matrix $\mathbf{\Sigma}_{WC}$ is considered as the prior and an estimate of the class covariance matrix is given by:

$$\check{\mathbf{\Sigma}}_j = \mathbf{\Sigma}_{WC} \frac{\tau}{\gamma_j + \tau} + \hat{\mathbf{\Sigma}}_j \frac{\gamma_j}{\gamma_j + \tau} \tag{2.2}$$

where τ is a control constant and γ_j is occupation count for class j. Obviously, if insufficient data is available for current class, the prior resource $\mathbf{\Sigma}_{WC}$ is considered as more reliable than the class estimation $\hat{\mathbf{\Sigma}}_j$. In case of infinite data, only the class estimation of covariance matrix $\hat{\mathbf{\Sigma}}_j$ is used for further processing.

2.2.3 Silence Reduction in HLDA estimation - SR-HLDA

From the point of view of transformation estimation, silence is a “bad” class as its distributions differ significantly from all speech classes. Moreover, training data (even if end-pointed) contains significant proportion of silence. An estimation of two HLDA transforms solves this problem but it makes the implementation more difficult.

Rather than discarding the silence frames, the occupation counts, γ_j, of silence classes ζ, which takes part in computation of HLDA estimation in equation are scaled by factor $1/SR$.

$$\hat{\gamma}_j = \frac{\gamma_j}{SR} \quad \text{if } j \in \zeta \tag{2.3}$$

$SR = \infty$ corresponds to complete elimination of silence statistics.

System	CTS system - WER [%]	Meeting system - WER [%]
(no HLDA)	36.71	30.3
standard HLDA	34.80	28.84
SHLDA	34.62	28.61
MAP-SHLDA	34.57	28.57
SR-HLDA	34.48	28.50

Table 2.1: Comparison of HLDA systems on eval01 and RT05 test sets.

2.3 Summary

Table 2.1 summarizes the performances of all already presented techniques. Smoothed HLDA (SHLDA) and MAP variant of SHLDA, taking into account the amounts of data available for estimation of statistics for different classes, perform both better than the basic HLDA. We have however found, that removing the silence class from HLDA statistics (Silence-reduced HLDA) is equally effective and cheaper in computation. Testing SHLDA and MAP-SHLDA on the top of SR-HLDA did not bring any further improvement, therefore we stick with SR-HLDA, especialy with complete silence removal, as the most suitable transformation in our next LVCSR experiments.

Chapter 3

Narrow band - wide band adaptation

As was already mentioned, 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. The common approach is to use other corpora for the training of acoustic models. 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). This data differs from the meeting domain, so one would try to adapt to either different recording environments or to different speech type. As the speaking style is often the cause for greater variability, adaptation to database with similar speaking style is generally preferred. Hence for the meeting domain, adaptation of models trained on CTS data is appropriate [2].

Conversational telephone speech speaking style matches well with meetings but CTS is naturally recorded with low bandwidth. Therefore, an adaptation to meeting domain is not trivial as the standard bandwidth for meeting recordings is 16 kHz (wide-band, WB).

3.1 Adaptation of CTS model to downsampled data (NB-NB)

The intuitive way to circumvent the problem of different band-widths is to downsample meeting data to NB and adapt CTS models into this domain, see Figure 3.1.

To find out the degradation caused by downsampling the data, the HMMs were

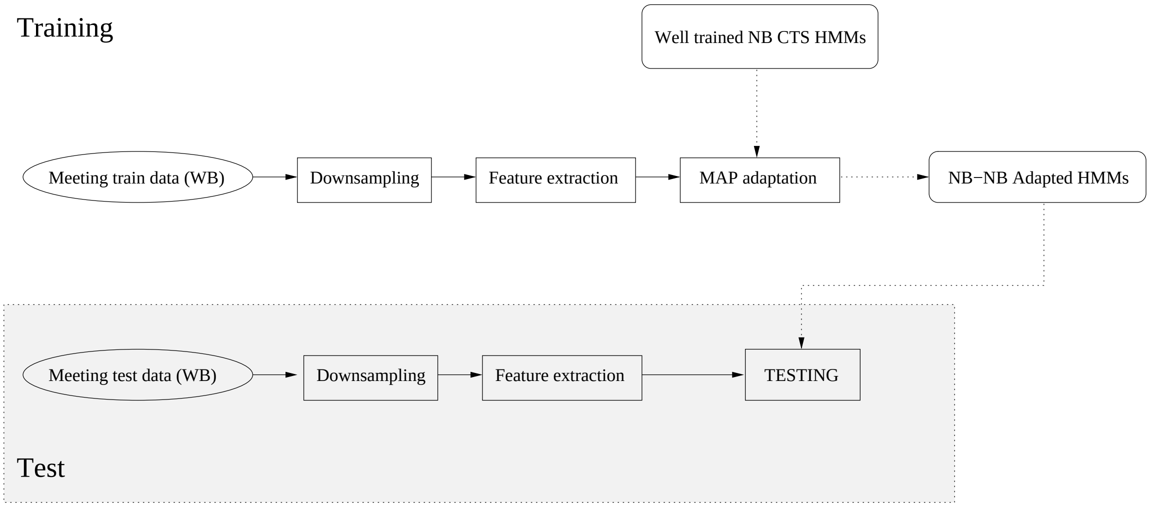

Figure 3.1: Simple system based on downsampling of WB data and adapted CTS models.

Training set	Adaptation	WER [%]
WB meeting	none	30.3
NB meeting	none	30.7
CTS	none	32.5
CTS-NB	MAP	29.8
CTS-NB	CMLLR MAP	29.8
CTS-NB	MLLR MAP	29.5

Table 3.1: Performance of non-adapted and downsampled systems.

also trained on downsampled meeting training data. The comparison is shown in the first two lines of Table 3.1. A degradation of 0.4% can be seen. The direct decoding of downsampled test data by CTS models does not perform well probably due to data mismatch - 2.2% worse than WB meeting system. But the adapted CTS system improves this result significantly. We tried to use just MAP adaptation (Figure 3.1) and also a cascade of MLLR or CMLLR followed by MAP adaptation, which outperforms WB baseline by 0.8% absolute.

The main **disadvantage** of this approach is the loss of the upper band (4-8 kHz) while it is known to contain useful information [5]. The solution will be given in the next section.

3.2 CMLLR as a transformation between wide-band and narrow-band

The loss of information by downsampling of WB data can be solved by global transformation based on Constrained Maximum Likelihood Linear Regression (CMLLR) to perform a WB to NB conversion and its application on WB features instead of downsampling the data. With this approach, even though the upper band information cannot be recovered, we can still make use of the richer information in actual target domain recordings. When the WB features are rotated to NB domain, MAP adaptation followed to settle the CTS models models into the target space.

In this approach, 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 the meeting data more like CTS data may however preserve some of the characteristics only visible with higher bandwidth. The basic idea of this process is shown in Figure 3.2.

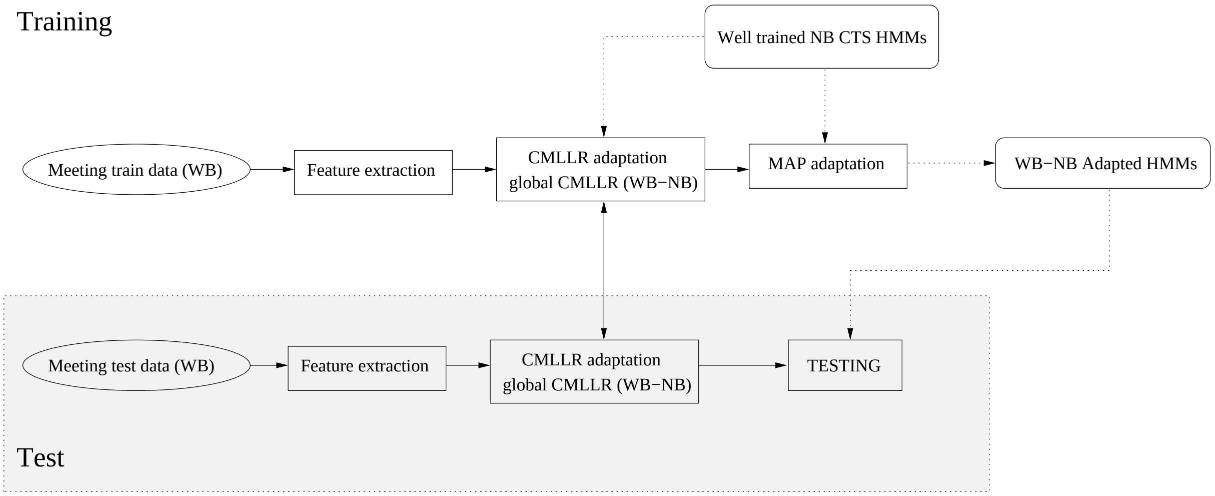

Figure 3.2: WB→NB adapted system based on CMLLR.

Training set	Adaptation	WER [%]
WB meeting	none	30.3
NB meeting	none	30.7
CTS-NB	CMLLR MAP	29.8
CTS-WB	CMLLR$_{WB\rightarrow NB}$,MAP	**29.1**

Table 3.2: Performance of WB→NB systems.

The accuracies of unadapted systems (training on the meeting data only), WB→NB adapted and downsampled systems are shown in Table 3.2. The best performance of 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.

3.3 WB→NB transform in HLDA estimation

3.3.1 WB→NB system based on HLDA from CTS

The easiest way to train WB→NB HLDA system is to take HLDA transforms and HMMs trained on CTS data and adapt them directly to the WB→NB transformed features similarly as in the system above. Obviously, this is not optimal as the HLDA is trained on CTS but the target data are meetings.

3.3.2 Adaptation of statistics

It is useful to estimate statistics from both data sets, to take an advantage of the meeting data also for HLDA matrix estimation. We use MAP adaptation of statistics, so 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.

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. This process is shown in Figure 3.3.

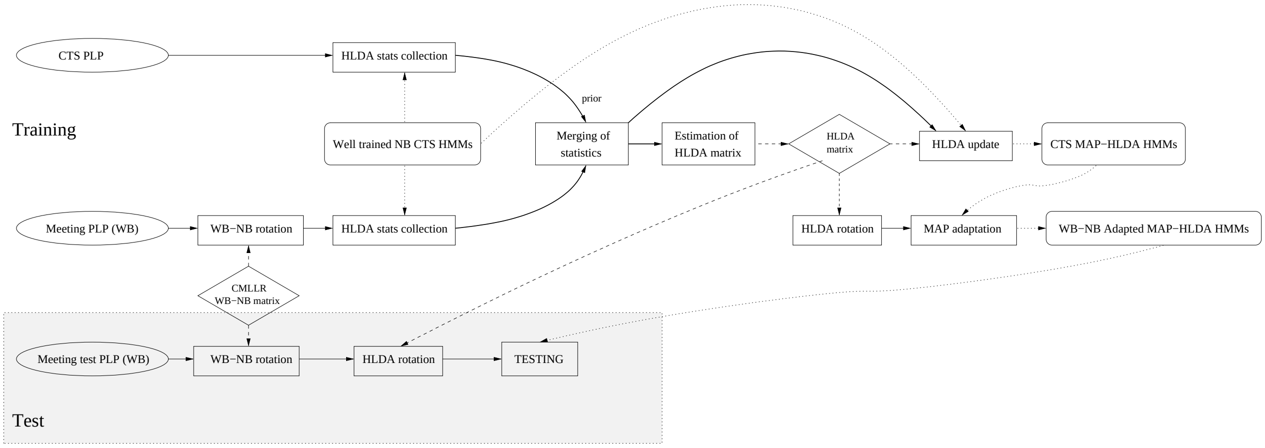

Figure 3.3: WB→NB system using HLDA based on merged statistics from the CTS and meeting training set.

System	HLDA adaptation	WER [%]
WB	none	28.5
NB	none	29.7
CTS-NB	MAP HLDA	29.0
WB-NB	CMLLR$_{WB\rightarrow NB}$, MAP HLDA	**27.8**

Table 3.3: Performance of HLDA systems.

3.3.3 Summary

For comparison with standard downsampling approach, the same experiments run also in the downsampled NB domain. First, NB non-adapted HLDA system was trained and evaluated. Table 3.3 shows 1.2% degradation of accuracy by downsampling.

Using an adapting scheme, the WB→NB CMLLR feature rotation in Figure 3.3 was replaced by downsampling of waveforms and feature extraction from this data.

3.4 WB→NB transform in Speaker Adaptive Training

Speaker adaptive training (SAT) was further used in addition to HLDA system to improve the accuracy. SAT is a technique used to suppress cross-speaker variance.

The comparison of proposed SAT implementations and traditional approach us-

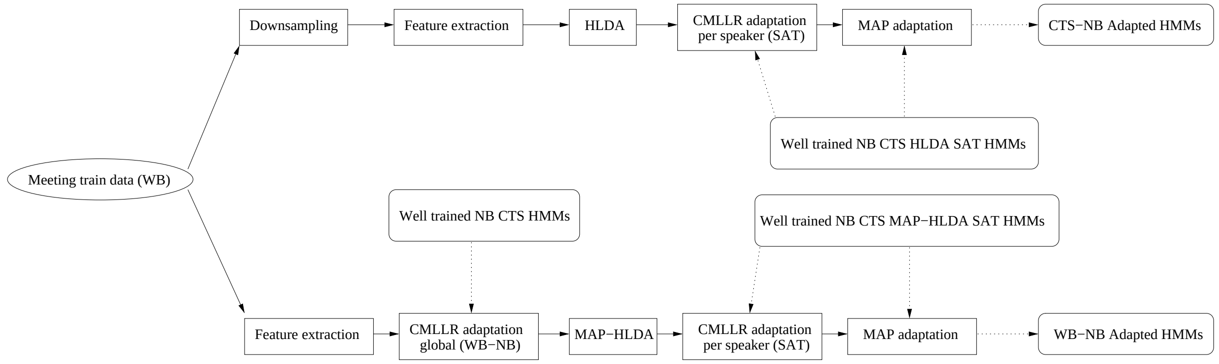

Figure 3.4: Downsampled and WB→NB adapted HLDA SAT system.

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$_{SAT}$	26.5

Table 3.4: Results of HLDA SAT systems

ing downsampling lies in replacing of WB→NB CMLLR by downsampling of data and feature extraction. It is shown in Figure 3.4. The upper branch presents the traditional approach and the lower branch the WB→NB system.

Table 3.4 shows the accuracy of the SAT systems. The best performance 26.6% is obtained by the WB→NB HLDA SAT system which is a 3.3% relative improvement over the non-adapted HLDA SAT system and 4.6% relative improvement over NB-NB adapted system.

3.5 Discriminative training of WB→NB adapted system

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 the best performance is achieved by using the Minimum Phone Error (MPE) criterion. MPE-MAP adaptation is an 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 the CTS prior to rotated adaptation data using iterative ML-MAP 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 a single iteration of ML-MAP adaptation, performing the iterative ML-MAP before starting the discriminative adaptation turned out to be essential for successful use of MPE-MAP.

3.5.1 Discriminative training of WB→NB adapted HLDA SAT system

A need for good CTS prior led to additional MPE training of the CTS_MAP-HLDA_SAT models. It produced new models set referred CTS_MAP-HLDA_SAT_MPE. When processing the meeting data, SAT transforms were estimated based on the CTS WB→NB resulting models from section 3.4. The transforms remained fixed for further processing.

To get good starting point for MPE-MAP adaptation CTS_MAP-HLDA_SAT models were adapted into WB→NB rotated domain using iterative ML-MAP with application of the above SAT transforms. These models, further referred to as NBWB_MPE_ML-MAP_SAT (see Figure 3.5).

Table 3.5 shows that a 1.5% absolute improvement is obtained by MPE training of ML-MAP adapted models. Incorporation of the CTS prior gives 0.3% additional improvement.

The final models were successfully used in the AMI LVCSR system for NIST

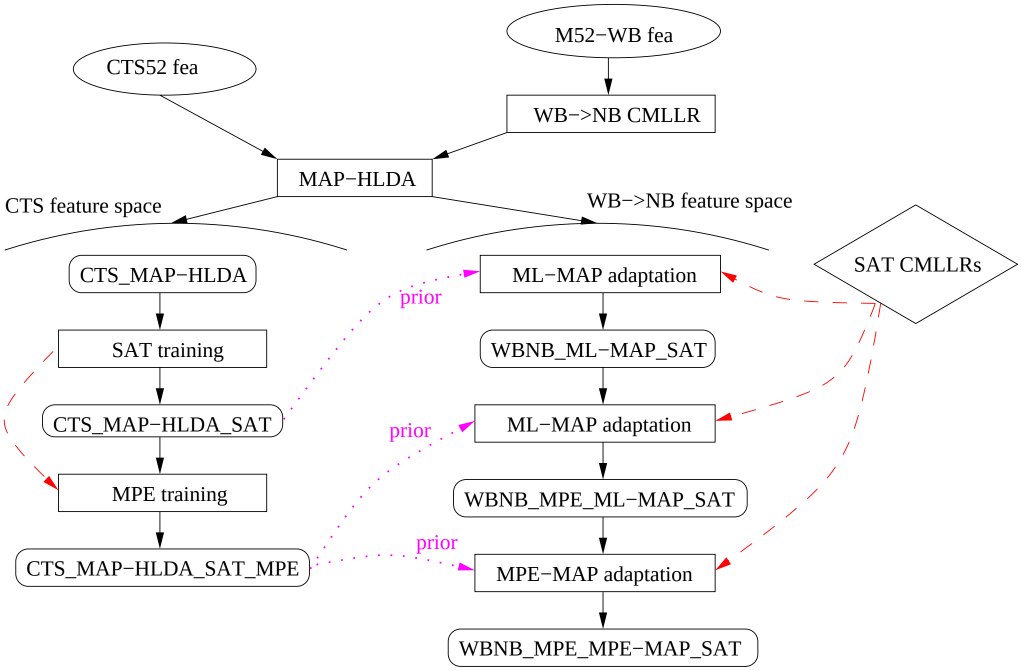

Figure 3.5: SAT - adaptation scheme of MPE-MAP adaptation into the WB→NB features.

Prior	Starting point	Adaptation	WER[%]
CTS_MAP-HLDA_SAT_MPE	-	ML-MAP	25.7
-	NBWB_MPE_ML-MAP_SAT	MPE	24.2
CTS_MAP-HLDA_SAT_MPE	NBWB_MPE_ML-MAP_SAT	MPE-MAP	**23.9**

Table 3.5: MPE-MAP in the SAT: Effect of prior and starting point.

2006 Rich Transcription evaluation.

3.6 WB→NB adapted system trained on boosted amount of data

The availability of new meeting data resources, especially a release of full AMI corpus[1], led to further improving of the current LVCSR system. Moreover, we were able to check how the techniques generalize on the new data. The CTS training data size was increased too by Fisher corpus. The new data sizes are described in Table 3.6. We see that the additional data represents about 70 hours of meetings and more than 1700 hours of telephone speech.

[1] Detailed information on AMI corpus is available at `http://corpus.amiproject.org`

Task	IHM	MDM	CTS
standard data	112h	63h	278h
boosted data	183h	127h	2000h

Table 3.6: Amount of data in the original and data boosted systems.

Data amount	112h	182h
ML HLDA SAT	27.5	25.8
MPE HLDA SAT	24.5	23.4

Table 3.7: Unadapted meeting system: Dependency of WER on the training data size.

All experiments to adapt a new 2000h CTS MPE HLDA SAT models into the WB→NB rotated domain used same algorithm as described above.

First, an unadapted baseline system was trained just on the new meeting data which yielded 1.7% absolute improvement in ML training over the original system and more than 1% when using MPE (see Table 3.7).

To capitalize on these gains, the 2000h CTS MPE HLDA SAT models were adapted in the WB→NB rotated domain according to the scheme in section 3.5.1: First, MPE starting point models were trained using ML-MAP and MPE-MAP adaptation followed.

Table 3.8 shows a 1.8% absolute gain due to adding training data and 1.3% improvement by adaptation from CTS.

Data amount	112h / 278h	183h / 2000h
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 3.8: WB → NB: Effect of training data and adaptation approach.

Chapter 4

Conclusions

The recognition of meeting speech is an important research issue and has been in the center of interest of several EC-sponsored projects: M4[1], CHIL[2], AMI[3], and AMIDA[4]. This work has been done in tight cooperation with the meeting recognition team in the series of M4/AMI/AMIDA and concentrated on feature extraction and acoustic modeling. It has investigated into two important problems in building of recognition system for meeting data:

- Improving of robustness of HLDA estimation by smoothing.
- Making use of additional data resources in the training.

4.1 Robust HLDA

Two approaches of HLDA smoothing were tested: Smoothed HLDA (SHLDA) and MAP variant of SHLDA, taking into account the amounts of data available for estimation of statistics for different classes. Both variants perform better than the basic HLDA. Moreover, we have found, that removing the silence class from the HLDA estimations (Silence-reduced HLDA) was equally effective and cheaper in computation. Testing SHLDA and MAP-SHLDA on the top of SR-HLDA did not bring any further improvement.

[1] `http://www.dcs.shef.ac.uk/spandh/projects/m4/`

[2] `http://chil.server.de`

[3] `http://www.amiproject.org/`

[4] `http://www.amidaproject.org/`

4.2 NB-WB adaptation

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 priors for adaptation. A solution of how to apply this transform for HLDA and SAT systems was given using maximum likelihood. 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.

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

4.3 Future work

In HLDA, the improvements obtained by smoothing techniques in HLDA show that these approaches are performing well but the differences are quite small. Therefore we want to focus on areas which suffer from higher insufficiency of data, such as long-span features [7], where the proposed approaches should lead to significant improvements.

In the second field, we have shown that speech data from sources different from the target domain can be advantageously used to improve the performance of a recognition system. The future of meeting recognition is definitely in processing speech from multiple multiple distant microphones (MDM), as they are much more practical for users than independent head-set microphones (IHM) which were used in this thesis. Therefore, our research will focus into this area. The potential for improvement is even greater than for IHM, as MDM speech corpora are even less available due to removing of crosstalks (more speakers talking in same time) and the channel variability in MDM speech is greater than for IHM.

Bibliography

[1] L. Burget. Combination of speech features using smoothed heteroscedastic linear discriminant analysis. In *8th International Conference on Spoken Language Processing*, Jeju Island, Korea, oct 2004.

[2] 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.

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

[4] J. Gauvain and C. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture. *IEEE Trans. Speech and Audio Processing*, 2:291–298, 1994.

[5] 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 *Proceedings of Interspeech 2005*, Lisabon, Portugal, 2005.

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

[7] Bing Zhang, Spyros Matsoukas, Jeff Ma, and Richard Schwartz. Long span features and minimum phoneme error heteroscedastic linear discriminant analysis. In *Proc. of DARPA EARS RT-04 workshop*, Palisades, NY, December 2004.

Author

Ing. Martin Karafiát
`http://www.fit.vutbr.cz/˜karafiat`

Martin Karafiát was born on November 6, 1976 in Pelhřimov, Czech Republic. He received his Master's degree in Electrical Engineering from the Brno University of Technology in June 2001. The topic of his diploma project was the online recognition of digit.

He joined the Laboratory of Signal Processing at Brno University of Technology as a Ph.D. student in September 2001. There he started to work on continuous speech recognition. He was on an internship in Speech and Hearing Group in University of Sheffield, United Kingdom from febuary 2003. Under a guidance of Prof. Steve Renals, he worked on speech recognition system for M4 project. He returned on December 2003 back to Brno and stayed continuing on M4 project.

In June 2004, he was again on 3 months internship in Speech and Hearing Group in University of Sheffield. Under a guidance of Thomas Hain he worked on recognition system for AMI project. After his return to Brno University in September 2004, he continued to work on AMI project.

During graduate studies, Martin Karafiát has authored and co-authored several conference papers presented on international events. He taught exercizes in signal and speech related courses and led several Ms. thesis. He is member of IEEE.

Abstract

This thesis investigates into two important issues of acoustic modeling for automatic speech recognition (ASR). The first topic are robust discriminative transforms in feature extraction. Two approaches of smoothing the popular Heteroscedastic Linear Discriminant Analysis (HLDA) were investigated: Smoothed HLDA (SHLDA) and Maximum A-Posteriori (MAP) adapted SHLDA. Both variants perform better than the basic HLDA. Moreover, we have found, that removing the silence class from the HLDA estimations (Silence-reduced HLDA) is equally effective and cheaper in computation. The second part deals with using heterogeneous data resources in ASR training. For a task, where little data is available for the target domain (meetings – 16kHz "wide-band" (WB) speech), techniques that allow to make use of abundant data from other domain, yet different in the acoustic channel (telephone data – 8kHz "narrow-band" – NB) were investigated. We successfully implemented an adaptation with WB data transformed to the NB domain based on Constrained Maximum Likelihood Linear Regression (CMLLR). A solution of how to apply this transform for HLDA and speaker-adaptive trained (SAT) systems was given using maximum likelihood. Finally, integration of this method with discriminative approaches was investigated and successfully solved. All experimental results are presented on standard data from NIST Rich Transcription (RT) 2005 evaluations.