INTEGRATION OF HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS (HLDA) INTO ADAPTIVE TRAINING

*Georg Stemmer**

Siemens AG, Corporate Technology
München, Germany
georg.stemmer@siemens.com

Fabio Brugnara

ITC-irst, Centro per la Ricerca Scientifica e Tecnologia
Povo (Trento), Italy
brugnara@itc.it

ABSTRACT

The paper investigates the integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into adaptively trained speech recognizers. Two different approaches are compared: the first is a variant of CMLLR-SAT, the second is based on our previously introduced method Constrained Maximum-Likelihood Speaker Normalization (CMLSN). For the latter both HLDA projection and speaker-specific transformations for normalization are estimated w. r. t. a set of simple target-models. It is investigated if additional robustness can be achieved by estimating HLDA on normalized data. Experimental results are provided for a broadcast news task and a collection of parliamentary speeches. We show that the proposed methods lead to relative reductions in word error rate (WER) of 8% over an adapted baseline system that already includes an HLDA transform. The best performance for both tasks is achieved for the algorithm that is based on CMLSN. When compared to the combination of HLDA and CMLLR-SAT, this method leads to a considerable reduction in computational effort and to a significantly lower WER.

1. INTRODUCTION

Heteroscedastic Linear Discriminant Analysis (HLDA) [1] is a linear transformation that performs both a diagonalization and a reduction of the dimension of the feature space. It may be regarded as a generalization of Linear Discriminant Analysis (LDA) which takes into account the individual covariance matrices of the classes. In this paper we investigate the integration of HLDA into an *adaptively trained* speech recognizer. Adaptive training schemes are often applied for tasks where recording conditions, acoustic environment or speaker frequently change. These algorithms estimate speaker-specific transformations –either of the models or of the acoustic features– to exclude phonetically irrelevant variability from training. The latter, which includes all kinds of speaker, channel and environment variability, is subsumed here by the single term *speaker variability*. Well-known representatives of adaptive training algorithms are *Speaker Adaptive Training (SAT)* [2] and *Vocal Tract Length Normalization (VTLN)* [3]. Classical SAT uses *Maximum Likelihood Linear Regression (MLLR)* [4] as a speaker-specific transformation, therefore it will be referred to as *MLLR-SAT* in this paper. Often *CMLLR-SAT* [5] is employed, a variant in which MLLR has been replaced by *Constrained Maximum Likelihood Linear Regression (CMLLR)* [6, 5]. CMLLR has the advantage that it allows for an efficient implementation as a feature transformation.

*The work was performed while the first author was at ITC-irst, Trento. It was partially financed by the European Commission under the project TC-STAR, contract reference number FB6-506738.

Integrating HLDA into an adaptive training procedure should ideally lead to a robust estimation both of the speaker-specific transformations and of the HLDA projection matrix. Furthermore the required computational resources (memory and time) should be as low as possible. A straightforward approach, which is investigated in this paper, is to integrate HLDA into CMLLR-SAT. An alternative is based on a normalization method for adaptive training that we have introduced recently: *Constrained Maximum-Likelihood Speaker Normalization (CMLSN)* [7, 8] is similar to VTLN in the sense that the speaker-specific transforms are applied before training the recognition-models, and it is similar to CMLLR-SAT in the sense that CMLLR is used for the feature transformation. We have shown [7, 8] that CMLSN leads to better results both than CMLLR-SAT and VTLN for a large number of different tasks. Thus, the integration of HLDA into CMLSN is of interest. It is shown for two different tasks that the proposed approach is effective; it even performs significantly better than the combination of CMLLR-SAT with HLDA. The procedure does not require a training of complex recognition-models in the high-dimensional feature-space which leads to a considerable reduction of computational effort. A third approach which is investigated in this paper is to evaluate if additional gains could be achieved by estimating HLDA in a speaker-normalized space. This way speaker-variability is reduced before optimizing the projection matrix and it can be estimated more robustly.

Previous work by Matsoukas and Schwartz [9] already investigated the problem to find a suitable integration of HLDA into adaptive training. Based on the assumption that speaker variability should be discarded from the HLDA projection, a variant of CMLLR which is based on full-covariance densities is estimated for each speaker in the high-dimensional original feature space. The projection matrix is then estimated from normalized features. The resulting *HLDA-SAT* algorithm requires to train and to store three different HMM sets for a speech recognizer. In the following we will investigate an approach to perform feature normalization before feature projection as well. However, the algorithm avoids the additional effort of full-covariance CMLLR by using semi-tied covariances. Furthermore only two different sets of HMMs are needed. In contrast to [9] our results do not confirm any significant advantage of the estimation of the HLDA projection in a normalized feature-space.

The paper is structured as follows: In Sec. 2 a short definition of HLDA is given. Sec. 3 is dedicated to an overview of CMLSN. In Sec. 4 different approaches to integrate HLDA into adaptive training are proposed and described. The data corpus and the baseline system are introduced in Sec. 5. Experiments are detailed and discussed in Sec. 6. We give a conclusion and a prospect on our future work in Sec. 7.

2. HLDA DEFINITION AND ESTIMATION

HLDA is defined in a maximum-likelihood framework [1]. The projection from a p-dimensional feature space into a q-dimensional feature space with $q < p$ is performed by a matrix $\mathbf{T}$ which is estimated iteratively with the EM-algorithm. Given matrix $\hat{\mathbf{T}}$ from the previous step and p-dimensional feature-vectors $\mathbf{O} := \mathbf{o}_1, \ldots, \mathbf{o}_T$ the improved matrix $\mathbf{T}$ is chosen such that

$$\mathcal{Q}(\hat{\mathbf{T}}, \mathbf{T}) = \sum_m \sum_t p(m|\mathbf{O}, \hat{\mathbf{T}}) \cdot \log [\, |\mathbf{T}| \cdot \mathcal{N}(\mathbf{T}\mathbf{o}_t | \mathbf{T}\mu_m, \mathbf{\Sigma}_m) \,]$$

is maximized. $\mathcal{N}(\mathbf{o}_t \mid \mu_m, \mathbf{\Sigma}_m)$ represents in this equation the output density m of an HMM. The covariance matrices $\mathbf{\Sigma}_m$ are constrained to be diagonal to allow for a row-by-row optimization of $\mathbf{T}$. $\mathbf{T}$ is optimized with an ad-hoc iterative numerical procedure [10] utilizing full-covariance statistics for each density m. In order to ensure that $\mathbf{T}$ is a projection for the optimization of the rows $q+1, \ldots p$ of $\mathbf{T}$ the state-specific statistics are replaced by the corresponding global covariance statistics. Consequently, after estimation of $\mathbf{T}$ the dimensions $q+1, \ldots, p$ of the transformed feature vectors $\mathbf{T}\mathbf{o}_t$ can be discarded, yielding q-dimensional feature vectors $\mathbf{o}_t^{\mathbf{T}}$. As HLDA performs a feature space diagonalization it is not reasonable to estimate the transformation matrix without jointly updating the covariance matrices of the output densities of the acoustic models $\mathbf{\Lambda}$. Thus, any algorithm for HLDA computation combines the optimization of the transformation with an update and a re-estimation of the covariances. The output is not only a matrix $\mathbf{T}$ but also an updated set of acoustic models $\mathbf{\Lambda}^{\mathbf{T}}$. We use an implementation of the time-efficient procedure introduced in [10], more details can be found in [11].

3. ADAPTIVE TRAINING USING CMLSN

Here we give a short review of the CMLSN algorithm for adaptive training. More details and a description of possible extensions can be found e. g. in [8]. The procedure is motivated by the aim of reducing the influence of speaker variability in the early stages of acoustic model training. CMLSN exploits the fact that CMLLR can be implemented as a feature transformation. Speaker-dependent CMLLR parameters $\mathbf{A}_s, \mathbf{b}_s$ are estimated using the EM-algorithm w. r. t. a model $\mathbf{\Lambda}$ from a set of feature vectors $\mathbf{O}_s$ for which a transcription has to be provided. In this work $\mathbf{O}_s$ does not necessarily correspond to a speaker's utterance but to a cluster of acoustically similar speech segments that have been determined in a data-driven manner. The transformed or normalized feature vectors $\mathbf{o}_{s,t}$ for a speaker s are computed according to $\mathbf{o}_{s,t} := \mathbf{A}_s\mathbf{o}_{s,t} + \mathbf{b}_s$. While CMLLR adaptation is usually performed w. r. t. the recognition-models, in CMLSN *separate* sets of models are used. Acoustic normalization is performed w. r. t. so-called *target-models* $\mathbf{\Lambda}^{\mathrm{n}}$. For the acoustic representation in the decoding phase *recognition-models* $\mathbf{\Lambda}^{\mathrm{r}}$ are employed. All parameters of the two model sets, like initialization, definition of the context-dependent allophones or model structure are completely independent. We have shown in [8] that it is advantageous to use target-models with a *simple* structure, i. e. triphone HMMs with a single Gaussian density for each tied-state. Adaptive training using CMLSN proceeds as follows:

1. train target-model $\mathbf{\Lambda}^{\mathrm{n}}$ on untransformed feature vectors $\mathbf{O}$.
2. for each speaker s, estimate $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{n}}$ for the feature vectors $\mathbf{O}_s$. Apply $\{\mathbf{A}_s, \mathbf{b}_s\}$ to $\mathbf{O}_s$, yielding transformed feature vectors $\mathbf{O}_s^{\mathrm{n}}$.
3. use the conventional training procedure to initialize and to train the recognition-models $\mathbf{\Lambda}^{\mathrm{r}}$ on $\mathbf{O}^{\mathrm{n}}$; including state tying and the definition of the context-dependent allophones.

In the recognition phase a transcription of the utterance has to be available from the first recognition pass. Normalization and decoding in the second pass is performed as follows:

1. estimate $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{n}}$ for the feature vectors $\mathbf{O}_s$. Apply $\{\mathbf{A}_s, \mathbf{b}_s\}$ to $\mathbf{O}_s$, yielding transformed feature vectors $\mathbf{O}_s^{\mathrm{n}}$.
2. decode $\mathbf{O}^{\mathrm{n}}$ using $\mathbf{\Lambda}^{\mathrm{r}}$.

In this paper CMLLR for feature normalization is combined with MLLR to adapt the recognition-models $\mathbf{\Lambda}^{\mathrm{r}}$ for the second recognition pass which gives usually a small additional improvement.

4. INTEGRATION OF HLDA INTO ADAPTIVE TRAINING

Three different approaches to integrate HLDA into adaptive training are compared in this paper. It has to be taken into account that CMLLR, which is used for feature normalization both in CMLLR-SAT and CMLSN, is based on a diagonal-covariance assumption. Consequently, it does not make sense to apply HLDA to a normalized feature vector and it is impossible to estimate HLDA directly for the recognition-models of an adaptively trained speech recognizer.

CMLLR-SAT+HLDA results from adding three iterations of CMLLR-SAT to HLDA-transformed recognition-models. It requires a significant amount of computational effort as the recognition models have to be estimated in the high-dimensional feature-space.

1. train recognition-models $\mathbf{\Lambda}^{\mathrm{r}}$ on high-dimensional features $\mathbf{O}$
2. estimate HLDA transform $\mathbf{T}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{r}}$, yielding transformed recognition-models $\mathbf{\Lambda}^{\mathrm{r},\mathbf{T}}$ and low-dimensional features $\mathbf{O}^{\mathbf{T}}$
3. Starting from $\mathbf{\Lambda}^{(0),\mathrm{r},\mathbf{T}} := \mathbf{\Lambda}^{\mathrm{r},\mathbf{T}}$ and features $\mathbf{O}_s^{(0)\mathrm{n},\mathbf{T}} := \mathbf{O}^{\mathbf{T}}$ iterate for $i = 1, 2, 3$:
   (a) estimate CMLLR $\{\mathbf{A}_s^{(i)}, \mathbf{b}_s^{(i)}\}$ w. r. t. recognition-models $\mathbf{\Lambda}^{(i-1),\mathrm{r},\mathbf{T}}$ and features $\mathbf{O}_s^{(i-1)\mathrm{n},\mathbf{T}}$ yielding normalized, low-dimensional features $\mathbf{O}_s^{(i)\mathrm{n},\mathbf{T}}$
   (b) perform one iteration of Baum-Welch-Training for $\mathbf{\Lambda}^{(i-1),\mathrm{r},\mathbf{T}}$ on $\mathbf{O}_s^{(i)\mathrm{n},\mathbf{T}}$, yielding updated recognition-models $\mathbf{\Lambda}^{(i),\mathrm{r},\mathbf{T}}$

In order to integrate the HLDA transform into the CMLSN algorithm it is estimated w. r. t. the target-models. This leads to a significant reduction in computation time and memory as only the target-models, which contain much less densities than the recognition-models, have to be estimated in the high-dimensional feature space. Two different methods are proposed as we want to evaluate if there is an advantage in reducing the speaker variability before collecting the statistics for HLDA estimation.

CMLSN+HLDA (A) applies HLDA and CMLLR sequentially, there is no reduction of speaker variability prior to HLDA. The dimension of the CMLLR transformation matrices is equivalent to the dimension q of the reduced feature space.

1. train target-models $\mathbf{\Lambda}^{\mathrm{n}}$ on high-dimensional features $\mathbf{O}$
2. estimate HLDA transform $\mathbf{T}$ w. r. t. the target-models, yielding transformed target-models $\mathbf{\Lambda}^{\mathrm{n},\mathbf{T}}$ and low-dimensional features $\mathbf{O}^{\mathbf{T}}$
3. estimate CMLLR $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. target-models $\mathbf{\Lambda}^{\mathrm{n},\mathbf{T}}$ and features $\mathbf{O}_s^{\mathbf{T}}$
4. train recognition-models on normalized, low-dimensional features $\mathbf{O}_s^{\mathrm{n},\mathbf{T}}$

CMLSN+HLDA (B) normalizes the feature vectors before application of HLDA. As conventional CMLLR is based on a diagonal-covariance assumption, the approach combines a *Semi-tied-Covariance transformation (STC)* $\mathbf{H}$ [10] with the estimation of the CMLLR transformations. This allows to compute a CMLLR for full-covariance matrices, with the constraint that the full-covariance-matrices have to be semi-tied ones. The dimension of the CMLLR transformation matrices is equivalent to the dimension p of the high-dimensional feature space before application of the HLDA transformation.

1. train target-models $\mathbf{\Lambda}^{\mathrm{n}}$ on high-dimensional features $\mathbf{O}$
2. estimate STC transform $\mathbf{H}$ w. r. t. the target-models $\mathbf{\Lambda}^{\mathrm{n}}$, yielding transformed target-models $\mathbf{\Lambda}^{\mathrm{n},\mathbf{H}}$ and diagonalized features $\mathbf{O}^{\mathbf{H}}$
3. estimate CMLLR $\{\mathbf{A}_s, \mathbf{b}_s\}$ w. r. t. $\mathbf{\Lambda}^{\mathrm{n},\mathbf{H}}$ and features $\mathbf{O}_s^{\mathbf{H}}$
4. map CMLLR and target-models back into original space: $\{\mathbf{A}_s, \mathbf{b}_s\} \mapsto \{\mathbf{H}^{-1}\mathbf{A}_s\,\mathbf{H},\ \ \mathbf{H}^{-1}\mathbf{b}_s\}$; $\mathbf{\Lambda}^{\mathrm{n},\mathbf{H}} \mapsto \mathbf{\Lambda}^{\mathrm{n}'}$
5. apply $\{\mathbf{H}^{-1}\mathbf{A}_s\,\mathbf{H},\ \ \mathbf{H}^{-1}\mathbf{b}_s\}$ to features $\mathbf{O}_s$, yielding normalized features $\mathbf{O}_s^{\mathrm{n}}$
6. estimate HLDA $\mathbf{T}$ w. r. t. target-models $\mathbf{\Lambda}^{\mathrm{n}'}$ and normalized features $\mathbf{O}_s^{\mathrm{n}}$, yielding low-dimensional features $\mathbf{O}_s^{\mathrm{n},\mathbf{T}}$
7. train recognition-models on normalized, low-dimensional features $\mathbf{O}_s^{\mathrm{n},\mathbf{T}}$

Note that the order of CMLLR and HLDA does also influence the estimation of the CMLLR matrices. HLDA is a transformation that reduces irrelevant variability in the projected feature space. Thus, we can suppose that CMLLR can be estimated more robustly in a HLDA-projected feature space. It has to be measured empirically which on of the methods (A) and (B) leads to better results in the final system. In Sec. 6 we will compare the algorithms (A) and (B) and we will relate them to CMLLR-SAT+HLDA.

5. DATA SETS AND BASELINE SYSTEMS

Experiments are conducted on a corpus of recorded parliamentary speeches and on a corpus of broadcast news data; in the following the tasks will be referred to as *EPPS English* and *HUB4*, respectively. *European Parliament Plenary Sessions (EPPS)* is one of the evaluation tasks of the EU-funded project TC-STAR (http://www.tc-star.org). The English EPPS training data, released within the TC-STAR project and consisting of about 40 hours of speech, are exploited for training. A trigram language model was trained on the EPPS English final transcriptions (about 36 million words) and then adapted to the manual transcriptions of the acoustic EPPS data (about 370,000 words). Results are reported for the EPPS English development test set which corresponds to about 3.5 hours of speech. For the HUB4 broadcast news task we used the BN-E data released by the LDC in 1997 and 1998 for training of the acoustic models. The corpora contain a total of about 143 hours of usable speech data. Language models were trained on ≈132 million words of broadcast news transcripts distributed by LDC and on the transcripts of the BN-E training data. For evaluation we use the 1998 Hub4 Eval98 data consisting of two files, each with 1.5 hours of speech. Results are reported w. r. t. the focus conditions *(F-conditions)* marked in this test set:

F0: baseline planned broadcast speech, clean background.
F1: spontaneous broadcast speech, clean background.
F2: speech over telephone, clean background.
F3: speech with background music.
F4: speech with degraded acoustics (noise, other speech).
F5: planned speech by non-native speakers, clean background.
FX: all other conditions that cannot be classified into F0-F5.

The front-end of the ITC-irst speech recognizer combines 13 Mel-frequency Cepstral Coefficients and their first- and second-order time-derivatives into a 39-dimensional feature vector. For the EPPS data manual segmentation is exploited while the HUB4 data is segmented using a Bayesian Information Criterion. For both tasks the speech segments are clustered automatically. In the baseline systems *Cluster-wise Mean and Variance Normalization (CMVN)* ensures that for each cluster the static features have mean zero and variance one. The acoustic models are state-tied, cross-word, gender-independent, bandwidth-independent triphone HMMs. The HUB4 baseline system has about 9000 tied states and about 146000 Gaussians while the EPPS baseline system has about 5000 tied states and about 91000 Gaussians.

6. EXPERIMENTAL RESULTS

Before integrating HLDA into adaptive training we were interested in measuring the performance of different popular adaptive training algorithms for a common test set. While we have already compared CMLSN with CMLLR-SAT in [8], we could not find any published comparison between MLLR-SAT and CMLLR-SAT. The results for the EPPS English task are shown in Tab. 1. Systems *MLLR-SAT (mean)*, *MLLR-SAT (mean+var.)* and *CMLLR-SAT* result from adding three iterations of SAT to the baseline system, they differ in the type of the speaker-specific transform employed for adaptive training: For *MLLR-SAT (mean)* and *MLLR-SAT (mean+var.)* 4-class MLLR adaptation is used while system *CMLLR-SAT* uses single-class CMLLR. In *MLLR-SAT (mean)* only the mean vectors are adapted in training, this corresponds to the definition of MLLR-SAT in [2]. For *MLLR-SAT (mean+var.)* both means and variances of the recognition-models are adapted in training. System *CMLSN* has been trained with the algorithm described in Sec. 3. The target-models are tied-state triphone-models with a single density per state. Note that for *CMLSN* only segment-wise mean normalization of the features is applied and no CMVN. All adapted systems in Tab. 1 use 4-class MLLR and the supervision of the baseline system in decoding. From Tab. 1 it can be seen that there is no advantage of adapting both means and variances in MLLR-SAT in contrast to mean-only adaptation. This may be due to the simplicity of the variance transformation, which is constrained to be a diagonal matrix for practical reasons. Furthermore, CMLLR-SAT and MLLR-SAT reach about the same performance which indicates that the specific type of the speaker-specific transform used in training does not have a large influence. As MLLR-SAT needs much more computational effort in training than all other approaches we will not take it into account in the following experiments. From all four adaptively trained speech recognizers, *CMLSN* performs best as it achieves a rel. reduction of

system	EPPS dev. set
baseline	16.1
baseline (adapted)	14.8
MLLR-SAT (mean)	14.2
MLLR-SAT (mean+var.)	14.2
CMLLR-SAT	14.1
CMLSN	13.6

Table 1. Word error rates (WER) on EPPS English for different adaptive training methods.

more than 8% over the adapted baseline and about 3% over *CMLLR-SAT*. This may be explained by the effective reduction of speaker variability already at the beginning of the training of the recognition-models in CMLSN and by the fact that the simple target-models which are used for estimating the CMLLR transforms have only a very limited ability to incorporate speaker variability. Next we analyze if these properties will still be advantageous when combined with HLDA.

All systems with HLDA investigated here are based on the same high-dimensional feature set that consists of the static features and their first, second and third-order time-derivatives, i. e. a total of $p = 52$ parameters. The dimension q of the projected lower-dimensional feature-space is 39. A comparison of the different approaches for the EPPS English task is shown in Tab. 2. The system

system	EPPS English dev. test set	
	w/o HLDA	HLDA
baseline	16.1	15.4
baseline (adapted)	14.8	14.3
CMLLR-SAT	14.1	13.7
CMLSN	13.6	13.2 (A) 13.1 (B)

Table 2. WER on EPPS English with and without HLDA.

configurations are the same as in Tab. 1. It can be seen that HLDA leads to a rel. reduction in WER between 3-4% for all systems. As the rel. improvement of *CMLSN* over *CMLLR-SAT* remains about the same also after application of HLDA we can conclude that the proposed methods (A) and (B) are effective. Obviously the projection matrix can be estimated robustly and reliably w. r. t. the target-models. This leads to a considerable reduction of computational effort as the target-models typically have much less output densities than the recognition models (here 5820 vs. 90287 densities). System *CMLSN+HLDA (A)* is about 8% better than the adapted baseline system with HLDA. However, as the difference between method (A) and (B) is negligible we cannot confirm our supposition that estimation of HLDA in a normalized feature space can improve results.

Our conclusions are affirmed by the results for the HUB4 task that are shown in Tab. 3. All adapted systems in this task use 2-class MLLR and the supervision of the baseline system. The application of HLDA leads to a rel. reduction in WER of about 6% for system *CMLSN*. The best system for the HUB4 task is *CMLSN+HLDA (A)* which is also the procedure that requires the lowest computational effort. The difference between *CMLLR-SAT+HLDA* and *CMLSN+HLDA (A)* is statistically significant according to the MAPSSWE test with a p-value of 0.003.

F-cond.	all	F0	F1	F2	F3	F4	F5	FX
proportion	100	30.7	19.3	3.4	4.3	28.2	0.7	13.5
baseline	20.5	12.9	20.0	30.0	24.0	20.7	20.9	34.3
baseline (adapted)	18.7	11.7	18.7	24.7	23.0	19.0	19.6	30.8
CMLLR-SAT	17.7	11.4	17.3	24.3	21.5	17.7	19.6	29.9
CMLSN	17.1	10.9	16.8	21.4	21.1	17.4	18.3	28.4
CMLLR-SAT+HLDA	16.5	10.1	16.0	23.4	21.7	16.4	17.0	28.4
CMLSN+HLDA (A)	16.1	10.1	15.9	22.9	19.8	15.8	17.4	27.8
CMLSN+HLDA (B)	16.4	10.6	15.9	22.7	21.1	16.2	17.0	27.7

Table 3. WER on HUB4 Eval98 for different approaches to integrate HLDA.

7. CONCLUSION AND FUTURE WORK

We have shown that HLDA can be effectively integrated into an adaptively trained speech recognition system. An evaluation of several algorithms demonstrated that HLDA can be estimated effectively w. r. t. to a set of simple target-models. When compared to an approach based on CMLLR-SAT the proposed method leads to a considerable reduction in computational effort for training and to significant reductions in WER. Our supposition that estimation of HLDA in a normalized feature space can improve results was not confirmed. Future work should include an experimental comparison of different input feature spaces for HLDA. Until now only third-order derivatives have been investigated. Furthermore, it seems promising to include scores of a Gaussian mixture model or of a phone-recognizer into the feature vectors before applying HLDA, similar to the approach described in [12].

8. REFERENCES

[1] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[2] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[3] L. Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP*, 1996, pp. 353–356.

[4] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[5] M.J.F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[6] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 5, pp. 357 – 366, 1995.

[7] D. Giuliani, M. Gerosa, and F. Brugnara, "Speaker normalization through constrained MLLR based transforms," in *Proc. ICSLP*, 2004, vol. 4, pp. 2893–2897.

[8] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in *Proc. ICASSP*, 2005, vol. 1, pp. 997–1000.

[9] S. Matsoukas and R. Schwartz, "Improved speaker adaptation using speaker dependent feature projections," in *IEEE ASRU Workshop*, 2003, pp. 273–278.

[10] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[11] G. Stemmer, "Heteroscedastic linear discriminant analysis (HLDA) – derivation, implementation and integration into adaptive training," Tech. Rep., ITC-irst, Centro per la Ricerca Scientifica e Tecnologia, 2005.

[12] G. Stemmer, V. Zeissler, C. Hacker, E. Nöth, and H. Niemann, "A phone recognizer helps to recognize words better," in *Proc. ICASSP*, 2003, vol. 1, pp. 736–739.