Fast unsupervised speaker adaptation through a discriminative Eigen-MLLR algorithm

Bart Bakker, Carsten Meyer, Xavier Aubert

Philips Research Laboratories, Weisshausstrasse 2, 52066 Aachen, Germany
{*Bart.Bakker, Carsten.Meyer, Xavier.Aubert*}*@philips.com*

Abstract

We present a new method for unsupervised, fast speaker adaptation that combines the Eigen-MLLR transform approach with discriminative MLLR. We thereby aim to profit both from the performance improvements that are generally provided by a discriminative approach, and from the reliability that Eigen-MLLR has demonstrated in fast adaptation scenarios. We present first evaluation results on the Spoke 4 subset of the 1994 Wall Street Journal (WSJ) database. Our results show that, in fast enrollment scenarios, discriminative Eigen-MLLR allows for clear improvements both over non-discriminative Eigen-MLLR and over discriminative MLLR. We further introduce a method to estimate the weight parameters of Eigen-MLLR discriminatively, and show that this allows for further improvements on the considered data sets.

1. Introduction

Fast, unsupervised adaptation is of the utmost importance for many practical applications of automatic speech recognition systems. Automatic telephone dialog systems and speech enabled user interfaces of consumer appliances, for example, rely on a swift adaptation to a random new speaker (preferably in the first few seconds). Such adaptation can be performed through MLLR or MAP algorithms, where a speaker-independent speech recognition system is transformed or partially re-optimised to better recognize an unknown new speaker. These methods do however perform poorly in on-line adaptation scenarios, where adaptation data is untranscribed and limited in length. Solutions to this problem have been found first in the Eigen-speaker approach [1], and later in Eigen-MLLR [2, 3]. The first relies on a series of full acoustic models (one per training speaker) that have been estimated *before* adaptation on a sufficiently large set of transcribed training data. Principal component analysis on these predefined models is used to construct a basis of 'Eigen-models' that has a much lower dimension than the full set of acoustic model parameters. During adaptation, the new acoustic model is expressed on this basis, i.e. estimated as a weighted sum of the predefined basis models. Eigen-MLLR follows the same strategy and constructs a new MLLR matrix as a weighted sum of pre-defined, Eigen-MLLR matrices. The prior information that is encoded in the Eigen-MLLR matrices allows a reliable estimation of the Eigen-weights, even when only limited and unsupervised adaptation data is available.

A second strategy to improve adaptation results can be found in the discriminative approach (see e.g. [4–8]). Discriminative training updates a speech decoder's acoustic parameters so that not only the target transcription becomes more likely to be recognised, but also makes erroneous decodings *less* likely. Discriminative methods have shown great success both in acoustic training and in adaptation scenarios. One of the most popular methods to perform discriminative training is maximum mutual information (MMI) estimation, which maximizes the mutual information between the acoustic signal and the target transcription, given the decoder's acoustic parameters. In this article we focus on an application of MMI to speaker adaptation, known as maximum mutual information linear regression (MMILR, see e.g. [7, 8]). Discriminative adaptation depends even stronger on the availability of sufficient transcribed adaptation data than MLLR adaptation, and by itself would appear challenging in a fast, unsupervised adaptation scenario.

We show however in this article that the strong points of Eigen-MLLR and discriminative adaptation (MMILR) can be successfully combined to create a powerful adaptation model that is still robust in fast adaptation scenarios. The key point is that the discriminative estimation of MLLR matrices is performed in the *training* phase, i.e. on annotated data (supervised estimation). The thus obtained MMILR matrices are better suited than MLLR to model each speaker's (speech) characteristics, and consequently provide a stronger basis for Eigen-adaptation. On-line adaptation subsequently requires only the estimation of the Eigen-weights, according to a suitable estimation criterion. Since the number of Eigen-weights is small to moderate, estimation can be performed in an unsupervised scenario as well, without a significant performance loss. Our method therefore allows for fast, unsupervised discriminative adaptation. The Eigen-weights can be obtained through an ML estimation, but also via a discriminative approach, which is introduced in this article.

The theoretical outline of the Eigen-MLLR method, discriminative estimation and its application to Eigen-MLLR are described in Sections 2 and 3. A more detailed mathematical description of the latter is given in the Appendix. Section 4 describes the experimental setup. Decoding results, expressed in word error rate, are presented in Section 5, where we show that unsupervised Eigen-MMILR outperforms both standard MMILR and Eigen-MLLR in fast enrollment scenarios. Section 6 concludes with a short discussion and outlook on future work.

2. Eigen-MLLR

Eigen-MLLR, as described e.g. in [2, 3], is an extension of standard MLLR where the transformation matrix for a new speaker is constructed as a weighted sum of predefined Eigen-MLLR matrices. These Eigen-matrices are estimated through a form of principal component analysis, which is implemented as follows:

Training. We estimate a series of speaker-dependent MLLR-matrices T_i ($i = 1..N_S$) on a set of N_S training speakers, for which a sufficient amount of *transcribed* speech data is avail-

able. We subtract the mean MLLR-matrix $\bar{T} = N_{\mathrm{S}}^{-1} \sum_{i=1}^{N_{\mathrm{S}}} T_i$ to yield transformation matrices $\tilde{T}_i = T_i - \bar{T}$. The speaker-dependent matrices $\tilde{T}_i$ are reshaped in the form of vectors $\tilde{\mathbf{t}}_i$, and we calculate the Eigen-values and Eigen-vectors of the corresponding covariance matrix $C = \sum_i \tilde{\mathbf{t}}_i \tilde{\mathbf{t}}_i^T$. Any new MLLR-matrix is now assumed to be of the form

$$T = \bar{T} + \sum_{i=1}^{N_e} \alpha_i \hat{T}_i \,, \tag{1}$$

where $\hat{T}_i$ denotes the i-th Eigen-vector of C (reshaped in the form of an MLLR-matrix). N_e is the number of Eigen-matrices that are actually used for adaptation (corresponding to the N_e largest Eigen-values), which may be much smaller than N_{S}, and α_i are the weight parameters that are to be estimated during adaptation. The Eigen-matrices can be constructed before adaptation, where we can make use of all the available training data.

Adaptation. We calculate the speaker-dependent weights through minimization of

$$E = \sum_{t=1}^{F} \sum_{s=1}^{S} \gamma_s(t) \left(o_t - \hat{\mu}(\alpha, s)\right)^T \Sigma_s^{-2} \left(o_t - \hat{\mu}(\alpha, s)\right) \,, \tag{2}$$

where F is the number of acoustic observation vectors o_t that are used for adaptation, S is the total number of states in the acoustic model, $\gamma_s(t)$ is the likelihood that observation vector t is decoded through state s, Σ_s is the variance for state s and

$$\hat{\mu}(\alpha, s) = \left[\bar{T} + \sum_{i=1}^{N_e} \alpha_i \hat{T}_i\right] \mu_s^{si} \,, \tag{3}$$

where μ_s^{si} is the unadapted (speaker independent) mean for state s. The number of weight parameters that are to be estimated generally is much lower than the number of elements in an MLLR-matrix. This makes Eigen-MLLR a robust method in scenarios such as fast, on-line adaptation, where no transcribed (and only limited untranscribed data) is available.

3. Discriminative estimation

Maximum likelihood (ML) estimation in speech recognition optimizes the match between the acoustic inputs that represent an audio signal and the acoustic model parameters that fit the corresponding transcription. In a stochastic framework, this means that the likelihood of the observed acoustic inputs given the transcribed text (or the recognised text in the unsupervised scenario) is maximized. Discriminative estimation, on the other hand, considers not only the correct classification for an acoustic training signal, but also the set of *in*correct classifications that have a high likelihood under the current model parameters. (Note that such classifications are not penalized in ML estimation.) To reduce the number of such misclassifications, each discriminative update changes the acoustic model parameters so that incorrect classifications become less likely, while maintaining a high likelihood for correct classifications. The difference between 'confusable' acoustic model parameters is increased in this manner, resulting in less 'confusion' between correct and incorrect classifications and in a lower error rate.

A well-known method to perform discriminative estimation is *Maximum Mutual Information Estimation* (MMIE). The maximum mutual information for one sequence of F frames reads

$$I = \log \frac{P(W, O_1^F|\theta)}{\sum_{W'} P(O_1^F|W', \theta) P(W')} - \log P(W) \,, \tag{4}$$

where $O_1^F = \{o_1, \ldots, o_F\}$ represents the sequence of acoustic observation vectors for the sentence, θ summarizes the (acoustic) model parameters and W represents the transcribed or the most likely word sequence. The second term is generally left out (since it does not depend on the acoustic model parameters), leaving the first term, which represents the logarithm of the *a posteriori* probability of the transcribed word sequence given the corresponding acoustic data. The sum over word sequences W' includes the set of misclassifications, which plays an important role in discriminative training. This set of word sequences is generally obtained through a free recognition on the training data. Both W and W' can be efficiently represented through a word or state lattice Λ (see e.g. [5]). We will refer to the lattices that represent W and W' as the 'positive' and the 'negative' lattice (Λ^+ and Λ^-), respectively. Both ML and discriminative training aim to increase the likelihood of the sequences in the positive lattice. Discriminative training (i.e. the maximisation of I), however, also aims to decrease the likelihood of the sequences in the negative lattice. Update formulas for discriminative training, for MMILR and for the weight parameters in Eigen-MMILR can all be obtained through maximisation of I with respect to the relevant model parameters. We elaborate on the corresponding mathematics in the Appendix.

4. Experimental setup

We apply our methods to the Spoke 4 part of the 1994 Wall Street Journal Database. This dataset contains read data from 4 American speakers. For each speaker an adaptation set of about 40 sentences (around 5 minutes), and a test set of the same size are available. Speaker independent model parameters were obtained from maximum likelihood training on the WSJ0 and WSJ1 sets. The 284 speakers (142 males and 142 females) in these sets were used to estimate speaker dependent *single class* MLLR and MMILR adaptation matrices to be used in the Eigen-scenarios. In the fast enrollment scenarios we used subsets of the adaptation data for adaptation, and the full test set for recognition. The used speech recognition system is based on 35-dimensional mel-cepstrum features, and uses continuous Laplacian mixture emission densities with a globally pooled variance vector. Recognition was performed with the WSJ 5K lexicon and the Baseline WSJ trigram language model.

Figure 1 describes the Eigen-adaptation scenario's. In the training phase, the speaker-dependent MLLR matrices are estimated via the following iteration: we use speaker independent ML parameters and the transcribed training data for speaker i to construct the positive state graph. This state graph is used to estimate the speaker specific transformation matrix $T_{i,1}$, and to obtain the updated acoustic parameters $\theta_{i,1}^{mllr}$. The latter are used to reconstruct Λ^+, and start the next iteration. After two such iterations the resulting MLLR matrices $T_{i,2}$ are used to construct the Eigen-MLLR basis, and the adapted model parameters $\theta_{i,2}^{mllr}$ are used as an initialization for MMILR adaptation. MMILR proceeds in the same fashion, only next to the positive state graph, we also construct the negative state graph corresponding to the incorrect or less likely word sequences. Both graphs are used to construct the Eigen-MMILR basis.

In the adaptation phase (either Eigen-MLLR or -MMILR adaptation), the weight parameters α_i are estimated, using the pre-constructed Eigen-MLLR or Eigen-MMILR basis matrices and unsupervised adaptation data. After two iterations, the resulting acoustic model parameters θ_2^{eigen} can be used for recognition. Discriminative estimation of the Eigen-weights is performed through an extra iteration, which is initialized with α_2^{eigen}

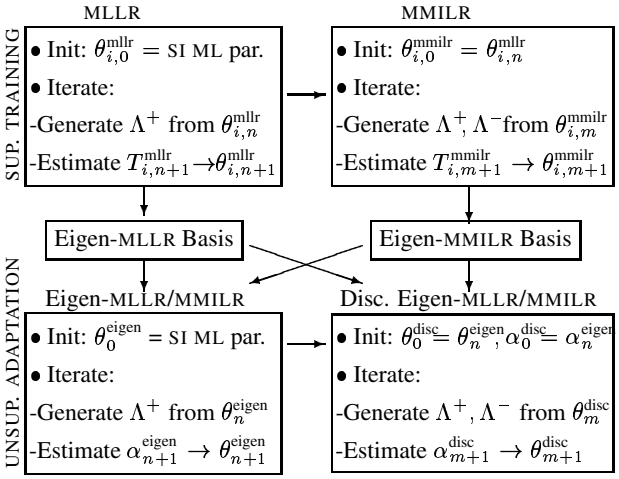

Figure 1: Flow diagram for the Eigen-adaptation methods that are used in this article. The top half describes the estimation of the Eigen-MLLR/MMILR basis matrices; the bottom half describes the estimation of the Eigen-weights, according to an ML criterion (left part), or through a discriminative approach (right part).

and θ_2^{eigen} (see Figure 1).

The iterations for standard supervised/unsupervised MLLR and MMILR also follow the outline indicated in the top half of Figure 1, only now based on adaptation data instead of training data.

5. Results

Table 1 contains the average word error rates over all speakers for the Spoke 4 set, for the 7 adaptation scenarios, and for the unadapted recogniser. For each speaker in the unsupervised scenarios we used 10 seconds of speech from the separate adaptation set for enrollment. Single class Eigen-MLLR and Eigen-MMILR adaptation methods were able to reduce the word error rate in the fast enrollment scenario where, as expected, the non-Eigen adaptation methods performed poorly and actually *increased* the WER[1]. It can further be seen that the improvement that is gained from the use of discriminative methods in the supervised scenario is present also in the unsupervised Eigen-adaptation scenario, where Eigen-MMILR clearly outperforms Eigen-MLLR. Discriminative estimation of the Eigen-weights in the Eigen-MMILR scenario produced a further improvement over the scenario where these weights were obtained from an ML estimation.

Figure 2 shows the influence of enrollment time (upper panel) and of the number of Eigen-matrices (N_e) that is used (lower panel) on the word error rate for unsupervised Eigen-MLLR and Eigen-MMILR. Note that Eigen-MMILR outperforms Eigen-MLLR at every instance. The lower panel shows that (for 10 seconds of enrollment data) the optimal number of Eigen-matrices is 50 for Eigen-MLLR adaptation, and 30 for the Eigen-MMILR method. These numbers of Eigen-matrices have been used in the experiments that correspond to the upper graph, both for Eigen-MLLR and Eigen-EMMILR adaptation. Note that the optimal numbers of Eigen-matrices are likely to be higher for longer enrollment times; this, however, has not been investigated in the present work.

[1] We observed that for adaptation times of more than one minute MLLR and MMILR produce better results than Eigen-adaptation, also in the unsupervised scenario. Such long enrollment times are however outside the focus of this article.

Enroll	Sup/Unsup	Method	WER	REL IMP
0 sec		No Adapt.	8.98%	
10 sec	unsupervised	MLLR	9.05%	-0.78%
10 sec	unsupervised	MMILR	9.43%	-5.01%
10 sec	unsupervised	EMLLR	8.36%	6.90%
10 sec	unsupervised	EMMILR	8.13%	9.47%
10 sec	unsupervised	DEMMILR	8.05%	10.4%
5 min	supervised	MLLR	7.60%	14.3%
5 min	supervised	MMILR	7.29%	18.8%

Table 1: Average word error rates over all test speakers, for the following scenarios: no adaptation, unsupervised MLLR and MMILR, unsupervised Eigen-MLLR and Eigen-MMILR (EMLLR and EMMILR), EMMILR with discriminative estimation of the weight parameters (DEMMILR), and supervised MLLR and MMILR. The relative improvement of the word error rate due to adaptation is also included for each adaptation scenario.

6. Discussion

In this article we have presented a new method for unsupervised adaptation, where we successfully combined the advantages of Eigen-MLLR and discriminative (MMILR) adaptation. Eigen-MLLR adaptation was able to reduce the word error rate by 6.9% (relative) in an unsupervised fast (10 sec) enrollment scenario, where non-Eigen adaptation methods performed poorly. This improvement was clearly surpassed by the Eigen-MMILR method which was introduced in this article, and yielded a relative WER decrease of 9.5%. Discriminative estimation of the Eigen-weights pushed this improvement still further, to 10.4% relative.

Eigen-MMILR adaptation relies on a basis of Eigen-MMILR matrices that is constructed from speaker specific, transcribed training data, and models the speech characteristics of the training speakers better than MLLR. We have shown that this *a priori* advantage holds in the on-line adaptation process, where only the weight parameters α_i need to be estimated: Eigen-MMILR outperformed Eigen-MLLR at every instance, including the fast (5-10 seconds) enrollment scenarios.

Discriminative adaptation of the weight parameters in the fast, unsupervised scenario requires the construction of a state lattice that contains all likely misclassifications which, in other scenarios, can be a time-costly exercise. The additional computation time in our scenario, however, is only limited: unsupervised adaptation already includes the construction of a word lattice (from a free recognition on the adaptation data) to find the most likely decodings; the 'less likely' part of the word graph can be used for discriminative estimation, at no extra cost. Furthermore, only few weight parameters are estimated on limited data, which means that the additional iteration can be performed in limited time as well.

More tuning of the speaker-dependent MMILR matrices may improve the Eigen-adaptation basis even further, and yield an extra WER reduction on the test data. Other extensions may include the use of multiple regression classes, MMI trained acoustic models as an initialization for MMILR adaptation and ML interpolation methods.

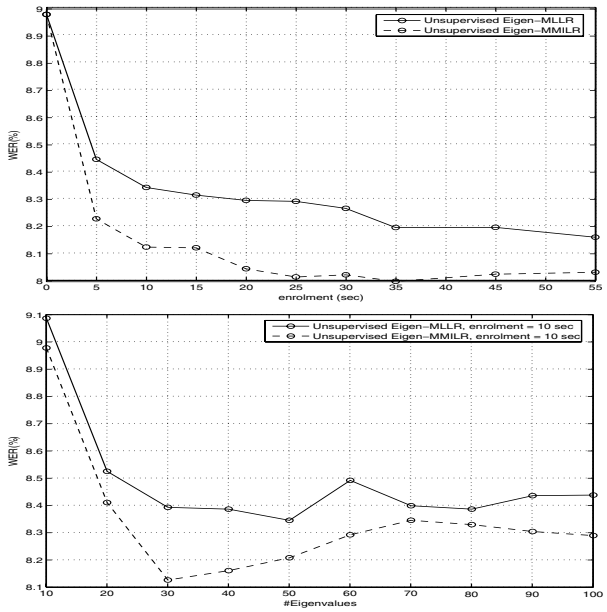

Figure 2: The word error rate for Eigen-MLLR (solid line) and Eigen-MMILR (dashed line), as a function of enrollment time (upper panel), and as a function of the number of used Eigenvalues (lower panel). For all graphs we used ML estimation of the Eigen-weights.

7. Acknowledgments

This research has been supported by a Marie Curie Fellowship of the European Community programme *Feedback Techniques in Automatic Speech Recognition* under contract number IST-2001-82941.

8. References

[1] Kuhn, R. *et al.* "Eigenvoices for speaker adaptation", Proceedings of ICSLP, vol. 5, pages 1771–1774, 1998.

[2] Chen, K. *et al* "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression", Proceedings of ICSLP, pages 742–745, 2000.

[3] Aubert, X. "Eigen-MLLRs applied to unsupervised speaker enrolment for large vocabulary continuous speech recognition", Proceedings of ICASSP-2004, pages 5–8, 2004.

[4] Normandin, Y. and Morgera, D. "An Improved MMIE Training Algorithm for Speaker-Independent Small Vocabulary, Continuous Speech Recognition", Proceedings of the IEEE, ICASSP-91, pages 537–540, 1991.

[5] Valtchev, V. *et al.* "Lattice-based discriminative training for large vocabulary speech recognition", Proceedings of the IEEE, ICASSP-96, pages 605–608, 1996.

[6] Woodland, P. and Povey, D. "Large scale discriminative training for speech recognition", Proceedings of the ASR-2000, pages 7–16, 2000.

[7] Uebel, L. and Woodland, P. "Improvements in linear transform based speaker adaptation", Proceedings of ICASSP-2001, pages 49–52, 2001.

[8] Gunawardana, A. and Byrne, W. "Discriminative speaker adaptation with conditional maximum likelihood linear regression", Proceedings of Eurospeech, 2001.

Appendix: Discriminative update formulas

In Section 3 we stated that all discriminative update formulas can be derived from the maximisation of I in Equation 4. Gunawardana and Byrne have demonstrated this for MMILR in [8], where (as an intermediate step) it is shown that that all discriminative update formulas can be found by solving the following equation:

$$\sum_{t=1}^{F}\sum_{s=1}^{S}(\gamma_s^+(t)-\gamma_s^-(t))\nabla_\theta \log P(o_t|s,\theta) \qquad (5)$$
$$+\sum_{t=1}^{F}\sum_{s=1}^{S} D_s \nabla_\theta \left\langle \log P(o_t|s,\theta)\right\rangle_{P(o_t|s,\theta')} = 0\ ,$$

where o_t is the acoustic feature vector at time t, ∇_θ is the derivative with respect to (any desired) free model parameter θ, θ' is the current best estimate of the model parameter and $\langle f(x)\rangle_{P(x)}$ denotes the expectation value of $f(x)$ under the probability distribution $P(x)$. The term D_s is a relaxation term which serves to stabilize the discriminative updates (see e.g. [8]). We further define

$$\gamma_s^- = P(o_t, s|O_1^F, \theta') \text{ and } \gamma_s^+(t) = P(o_t, s|W, O_1^F, \theta') \quad (6)$$

the probability that o_t is decoded through state s given all other acoustic outputs (γ_s^-) or given all other acoustic outputs *and* the transcribed or most likely word sequence W (γ_s^+). These marginal probabilities are generally expressed as the edge scores of the negative and the positive state graph. Note that the positive state graph is a subset of the negative state graph.

Discriminative updates for the MLLR matrices can be found by setting $\nabla_\theta = \nabla_{T_i}$ and inserting Gaussian densities $P(o|s) = N(o|\mu_s, \Sigma_s)$ with $\mu_s = T_i\mu_s^{si}$ in Equation 5 (see [7, 8] for the explicit formulas).

The discriminative update formula for the weight parameters in the Eigen-MLLR method follows in a similar fashion, when we insert Equation 3 into (5) and take $\nabla_\theta = \nabla_{\alpha_i}$, leaving

$$\sum_{t,s}\Bigg[(\gamma_s^+(t)-\gamma_s^-(t))\Big(o_t-(\bar{T}+\sum_{j=1}^{N_e}\alpha_j\hat{T}_j)\mu_s^{si}\Big)^T\Sigma_s^{-2}\hat{T}_i\mu_s^{si}$$
$$+D_s\Bigg(\mu_s^{si\,T}(\bar{T}+\sum_{k=1}^{N_e}\alpha'_k\hat{T}_k)^T\Sigma_s^{-2}\hat{T}_i\mu_s^{si}-$$
$$\mu_s^{si\,T}\hat{T}_i^T\Sigma_s^{-2}(\bar{T}+\sum_{j=1}^{N_e}\alpha_j\hat{T}_j)\mu_s^{si}\Bigg)\Bigg] = 0\ , \quad (7)$$

which can be rewritten as a set of N_e linear equations, each of the form

$$\sum_{j,t,s}\Bigg(\Big(\gamma_s^+(t)-\gamma_s^-(t)\Big)\mu_s^{si\,T}\hat{T}_i^T\Sigma_s^{-2}\hat{T}_j\mu_s^{si}+$$
$$D_s\mu_s^{si\,T}\hat{T}_i^T\Sigma_s^{-2}\hat{T}_j\mu_s^{si}\Bigg)\alpha_j =$$
$$\sum_{t,s}\Bigg(\Big(\gamma_s^+(t)-\gamma_s^-(t)\Big)(o_t-\bar{T}\mu_s^{si})^T\Sigma_s^{-2}\hat{T}_i\mu_s^{si}$$
$$+D_s\mu_s^{si\,T}\sum_k\hat{T}_k^T\alpha'_k\Sigma_s^{-2}\hat{T}_i\mu_s^{si}\Bigg)\ , \quad (8)$$

from which the weight parameters α_j can be solved.