Scaled Likelihood Linear Regression for Hidden Markov Model Adaptation

Frank Wallhoff, Daniel Willett†, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
{wallhoff,willett,rigoll}@fb9-ti.uni-duisburg.de

Abstract

In the context of continuous Hidden Markov Model (HMM) based speech-recognition, linear regression approaches have become popular to adapt the acoustic models to the specific speaker’s characteristics. The well known Maximum Likelihood Linear Regression (MLLR) [1] and Maximum A Posteriori Linear Regression (MAPLR) [2] are just two of them, which differ primarily in the training objective they are maximizing.

However, besides the approaches mentioned above there exists another known training objective which is the Maximum Mutual Information (MMI). By combining this MMI-approach with the linear regression of the HMM’s mean values, our research group developed a new adaptation technique that we call Scaled Likelihood Linear Regression (SLLR) as introduced in [3]. In this approach, the distance of the correct model sequence against the wrong ones is discriminated framewise. Like all techniques using MMI objectives, this adaptation is computationally very expensive compared to techniques using ordinary ML based objectives.

This paper therefore addresses the problem of an appropriate approximation technique to speed up this adaptation approach, by pruning the computation for tiny values in the discrimination objective. To further explore the potential of this adaptation technique and its approximation, the performance is measured on the LVCSR-system DUDeutsch developed by our research group at the Duisburg University and additionally on the 1993 WSJ adaptation tests of native and non-native speakers for the supervised case.

1. Introduction

Speaker adaptation has become an important sub-area in the field of Large Vocabulary Continuous Speech Recognition (LVCSR). Especially in the context of continuous HMMs, linear regression techniques have emerged to be an effective method to adapt the acoustic models to the individual characteristics of a certain speaker in order to improve recognition performance. In the past the well known MLLR [1] has become one of the most relevant techniques for speaker adaptation. The main idea of this technique is to adapt the mean values of the correct HMM-models, so that the Maximum Likelihood (ML) criterion is optimized.

However, in the past it has also shown that discriminative objectives like the maximum mutual information (MMI) showed their superiority over ML based approaches, in particular when the quantity of available training or adaptation material is low, as it is in cases for online adaptation tasks. Since it is known that discriminative objectives are computationally very expensive compared to ML-based ones, the main idea of this paper is to explore the influence of a pruning of those computations under a minimum probability threshold for the SLLR-adaptation. Therefore we decided to evaluate the performance of the SLLR adaptation on our German LVCSR-system called DUDeutsch (DD) and a part of the 1993 WSJ adaptation test.

In the next section, the speaker adaptation using linear regression is summarized first. After this the training objectives for MLLR and SLLR are revisited, followed by the introduction of the pruning approach for the SLLR. Then the recognition systems used are briefly presented. Before the paper closes with the conclusion and outlook, the experiments and achieved results are described.

2. Adaptation Using Linear Regression

The common basic idea of the adaptation techniques MLLR, MAPLR and SLLR is the unified linear transformation of the parameters of a large cluster of HMM states. In continuous HMM systems, the transformation is restricted to the Gaussians’ mean vectors \( \mu \), as they define the distributions’ major characteristics.

\[
\mu^* = M\mu
\]

By only estimating the transformation matrix or matrices \( M \) in the transformation formula above to the given adaptation data, the number of parameters that have to be estimated is very limited, which results in a robust estimation. The adjustment of the numbers of matrices to be estimated to the amount of available training data has become a common practice.

To determine the regression classes depending on the amount and type of available data we chose a procedure similar to the BIC-based procedure in [4]. When using several regression matrices, our procedure clusters the HMM states in a greedy way. Starting from initially unclustered HMM states always those two states or state clusters are chosen for being clustered that are the closest with respect to their mean vectors until all matrices have at least a given number of observations.

In order to enable the transformation not only to perform a rotation and a scaling but also a translation, the original mean vectors \( \mu \) should be regarded as expanded by one bias component of fixed value and thus \( M \) as being a \((d + 1) \times d\) matrix, where \( d \) represents the dimension of the feature vectors. The use of full transformation matrices seems to be superior as was proven by initial results in [5].

2.1. Maximum Likelihood Linear Regression

As introduced above, there are several optimization criteria for the estimation of the matrix or matrices \( M \). The most popular
approach is the ML criterion. In the following, we derive the presented adaptation data as a sequence of feature vectors \( X = (x_1, \ldots, x_T) \) which are labeled by a sequence of words \( W = (w_1, \ldots, w_T) \).

For this approach, the estimation of the free parameters of the acoustic models \( \lambda \) are estimated to maximize the likelihood of the adaptation data given by:

\[
\lambda_{\text{ml}} = \arg \max_{\lambda} P(X|W, \lambda)
\]  

(2)

In the case of a MLLR adaptation with a single regression matrix \( M \) that transforms all the Gaussians’ means, this yields

\[
M_{\text{mllr}} = \arg \max_{M} P(X|W, M)
\]  

(3)

Although the standard MLLR approach uses the EM-algorithm for the estimation of the regression matrix or matrices \( M \) (see [1]), we chose to use a gradient descent procedure, namely RPROP [6] to cope with this task. This does not make much of a difference except that the extension to a discriminative training objective as explained in the next section is more straightforward.

In the implementation that we chose for performing MLLR, we restricted training to the best path (Viterbi path). In this case, the regression matrix \( M_{\text{mllr}} \) is set up according to

\[
M_{\text{mllr}} = \arg \max_{M} \prod_{t=1}^{T} p(x_t|q_t, M)
\]  

(4)

where \( q_t \) represents the HMM state assigned to the feature vector \( x_t \) at time-frame \( t \) on the Viterbi path and \( p(x_t|q_t, M) \) represents the probability distribution function of state \( q_t \) with the mean vectors transformed by regression matrix \( M \). With \( p(x_t|q_t) \) modeled as a mixture of \( G_q \) with \( c_i \) weighted Gaussian distributions, and with \( M \) transforming the Gaussian means \( \mu_{q_t} \) 

\[
M_{\text{mllr}} = \arg \max_{M} \prod_{t=1}^{T} \sum_{i=1}^{G_q} \{ \frac{c_i}{\sqrt{2\pi} \sqrt{\Sigma_{q_t}}} \exp \left\{ -\frac{1}{2} \left[ \frac{x_t - \mu_{q_t}}{\Sigma_{q_t}} \right]^T \right\}
\]  

(5)

When using several regression classes \( C_1 \ldots C_n \) with independent regression matrices \( M_1 \ldots M_n \) and a function \( c(q) \) calculating the regression class number of state \( q \), equation 4 becomes

\[
M_{\text{mllr}}^1 \ldots M_{\text{mllr}}^n = \arg \max_{M_1 \ldots M_n} \prod_{t=1}^{T} p(x_t|q_t, M_{c(q_t)})
\]  

(6)

The section on experimental results shows the gain we are able to achieve when performing this type of MLLR on the DD-system and the 1993 WSJ adaptation test sets for native and non-native speakers. Because of the low computational costs of this criteria, there is no need for a further acceleration.

2.2. Scaled Likelihood Linear Regression

In numerous publications discriminative objectives like MCE or MMI [7, 8] proved their superiority over ML-based training objectives. Especially when training data and model resolution is very limited. It has turned out to be important not just to try to model the data as accurately as possible, but also to maximize the discrimination, because this has a higher influence to the recognition performance. This is due to the fact that for good recognition performance not the absolute likelihood of the correct word sequences but their likelihood difference against possible wrong sequences is of highest importance.

The main characteristic of discriminative training objectives is not only to maximize the observations’ conditional likelihood \( P(X|W, \lambda) \) given the true transcription \( W \), but to minimize its unconditional likelihood \( P(X|\lambda) \) at the same time. Thus, discriminative objectives try to maximize a term like

\[
\lambda_{\text{disc}} = \arg \max_{\lambda} \frac{P(X|W, \lambda)}{P(X|\lambda)}
\]  

(7)

In the equation above \( \lambda \) once again represents the set of parameters to be estimated.

A simple approach that is only very slightly affected by the distribution of confusion on the adaptation data is the frame-based discriminative training [9, 10], where \( P(X|\lambda) \) is approximated as

\[
P(X|\lambda) \approx \prod_{t=1}^{T} p(x_t|\lambda)
\]  

(8)

with \( p(x_t|\lambda) \) usually being further approximated according to

\[
p(x_t|\lambda) \approx \sum_{q \in Q} P(q|\lambda)p(x_t|q, \lambda)
\]  

(9)

as the average likelihood over all the states of the HMM system. The states’ prior probabilities \( P(q|\lambda) \) can simply be computed using the their frequency of occurrence within the training data.

By combining the equations 7, 8 and 9 the exhaustive frame-based discriminative objective can be formulated to

\[
M_{\text{sllr}} = \arg \max_{M} \prod_{t=1}^{T} \frac{p(x_t|q_t, M)}{\sum_{q \in Q} P(q)p(x_t|q, M)}
\]  

(10)

for a single regression class. And for multiple regression classes with independent matrices \( M_1 \ldots M_n \), this yields:

\[
M_{\text{sllr}}^1 \ldots M_{\text{sllr}}^n = \arg \max_{M_1 \ldots M_n} \prod_{t=1}^{T} \frac{p(x_t|q_t, M_{c(q_t)})}{\sum_{q \in Q} P(q)p(x_t|q, M_{c(q_t)})}
\]  

(11)

2.3. SLLR-Pruning

As the comparison of equations 5 and 10 directly shows, the number of computation steps for the SLLR vastly rises from \( T \) to \( T \cdot Q \), where \( T \) denotes the number of available adaptation frames and \( Q \) the amount of all models and all states on the Viterbi path.

The ability of an acceleration method is based on pruning. In the following just the computation of those derivatives for the regression matrices are taken into account of which the probabilities are most competing to correct model. Therefore equation 10 is further approximated by simply leaving out the computation of those derivatives, where the value of the products in the denominator drop under a certain given threshold \( \beta \). For each of the skipped frames applies:

\[
P(q) \cdot p(x_t|q, M_{c(q)}) < \beta
\]  

(12)

This means that for each skipped frame, the number of all computations to be made can be reduced by \((d + 1) \times d\) when using a fully occupied matrix \( M \). The value of \( d \) once again denotes the dimension of the observation-vector \( \vec{x} \). In the following just the (negative) exponent \( \kappa \) in the equation

\[
\beta = 10^{-\kappa}
\]  

(13)
is referred as the pruning-parameter, thus a pruning value of \( \kappa = 40 \) means that all frames with an average likelihoods greater than \( \beta = 10^{-40} \) are taken into consideration.

The resulting speedup \( \eta \) is defined by the equation below.

\[
\eta = \frac{\text{computation time of a full SLLR}}{\text{computation time of a pruned SLLR}}
\]

3. Description of the Deployed Systems

Using the adaptation techniques explained above, several experiments and their recognition rates are performed on the systems we call DD (DUDeutsch from Duisburg University Deutsch) and WSJ (Wall Street Journal). The systems used are described in more detail first.

The DD system is a LVCSR-system for German. It was developed by our research group and makes use of 100,000 words and a trigram language model, that was built out of data gathered in the internet. The speaker-independent acoustic models were previously trained using the German databases VERB-MOBIL and PHONDAT. The baseline WSJ-system was trained with about 7,200 utterances of different speakers, that are part of the 1992'Er WSJ-database set "Si-84". Furthermore, we use a standard bigram language model for the WSJ system with the number of 5,000 entries.

Both systems make use of tree-based clustered word-internal 3-state triphones (separately trained for both systems), modeled as Gaussian mixtures with of up to 20 components for the DD-system and with 10 for the WSJ-system. In both systems the dimension \( d \) of the acoustic feature vectors is 39 (12 MFCCs, logPower. \( \Delta d \) and \( \Delta \Delta d \)).

The performance of the adapted DD-systems are evaluated with 50 self recorded test utterances for three speakers. For the WSJ systems we had 10 native speaker from the 1993'Er adaptation test set S3C2 and 10 non-natives from the test set S3P0. The number of test utterances varied from 20 to 40.

The runtimes for the supervised adaptations with several parameter constellations and different numbers of adaptation utterances are measured on a Linux-workstation equipped with a Pentium III running at 850 MHz that are given in the tables. All experiments are supervised.

4. Experiments and Results

The first experiments are applied to verify the superiority of the frame discriminative objective over the maximum likelihood approach starting from the speaker independent models using the DD-system. In this experiments multiclass MLLR and SLLR-adaptations are performed using a different amount of available training material (10/50 utterances \( \approx \) 30/200 seconds). The groups of triphone models to be transformed simultaneously are clustered in that way, that there are at least 1250 observations for one resulting regression cluster.

Additionally we performed an adaptation using 150 utterances (uts) in three adaptation stages, which is labeled as MULTI in table 1. The first adaptation step consists of 100 iterations using one single regression matrix, followed by another 100 MLLR-iterations using multiple regression classes with corresponding matrices that have a minimum of 1250 observation-examples. In the third stage 20 SLLR-iterations with a pruning factor of \( \kappa=38 \) and a single regression matrix are performed.

The rows in the tables denoted by \( \Delta \text{abs} \) and \( \Delta \text{rel} \) contain the absolute and relative improvement of the average word error (Avg-WER) compared to the baseline system (BASE) in percent. The average computation time per iteration in minutes is given by \( \text{Avg-CT/it} \).

The influence of the pruning parameter, respectively \( \kappa \) on the recognition score and the speedup is evaluated in the following experiments using 10 iterations steps with 5 given utterances (\( \approx \) 15 seconds). As listed in Table 2, we varied \( \kappa \) from 30 up to 40. Additionally a full SLLR and MLLR adaptation are performed. In another experiment we ran 100 iteration steps of a pruned SLLR labeled as SLLR\( ^{\kappa} \) in the table below.

![Figure 1: Further adaptations with the DD system](image)

The left part of the figure above shows the relative reduction of the WER for a varying pruning parameter \( \kappa \) from 30 to 40 for a constant amount of adaptation material as well as an exhaustive SLLR and a MLLR. Further the relative improvement of the WER for a pruned SLLR (\( \kappa=40 \)), exhaustive SLLR and ordinary MLLR is depicted on the right part in the image for the number of 5, 10, 20 and 50 utterances.

As can be concluded from the tables and the figure above, the pruned and full SLLR adaptations achieve in almost all scenarios higher gains in the reduction of the WER than the corre-
sponding MLLR adaptations. We measured speedup values of nearly up to 10 for the pruned SLLR.

Beneath the adaptation experiments in conjunction with the DD system, further experiments with the better known WSJ database are performed. As above, we start with the evaluation of the baseline performance for the speaker independent models separated by natives and non-natives speakers as shown in the tables 3 and 4. After this, a combined adaptation similar to the one presented in [3] was performed that is used as the new baseline system for the further experiments. Therefore, the models were adapted in 100 iterations with one regression class and after this with another 100 iterations but several regression classes. The scores are presented with the label MULTI MLLR.

The following experiments for a pruned and a full SLLR as well as the MLLR adaptation for the native and non-native speakers refer to the already adapted systems and not to the speaker independent baseline system.

Again, these experiments and results show the superiority of the SLLR over the MLLR approach for the exhaustive discrimination as well as for a pruned one. The pruned adaptation that still achieves further reductions on the WER uses roughly just a fourth of the computation time of the exhaustive SLLR would need.

The results from all experiments above show, that the use of the introduced pruning technique achieves speedups of up to 10 for a pruned SLLR compared to an exhaustive one while simultaneously reducing the WER. It further turns out that further improvements concerning the WER can be achieved by this approach, especially for few adaptation utterances.

5. Conclusion and Outlook

The paper has introduced the pruned frame-discriminative adaptation technique SLLR for HMM based acoustic model adaptation. With this technique several experiments and tests were performed using the systems DD and WSJ with different parameter constellations such as quantity of the adaptation utterances and the value of the pruning factor. Once again it turned out that the SLLR adaptation outperforms the MLLR approach in several scenarios for both systems.

We achieved speedups concerning the computation time of up to 10 for the SLLR. Furthermore it has turned out that especially for few adaptation data, just the discrimination of the correct against the most competing models, with the product of the prior and the production probabilities over a certain threshold are of highest importance for the reduction of the WER.

In the future unsupervised online adaptation tasks seem to be a promising task in combination with the pruned SLLR technique and confidence measures as used in [11].

6. References