Some issues on the Expectation-Maximisation process for Maximum Likelihood Linear Regression

Míriam Luján-Mares, Carlos-D. Martínez-Hinarejos, Vicente Alabau, Alberto Sanchis

Institut Tecnològic d'Informàtica, Universitat Politècnica de València
Camí de Vera, s/n. 46071 València, Spain
{mlujan, cmartine, valabau, josanna}@dsic.upv.es

Abstract

The Maximum Likelihood Linear Regression (MLLR) technique has commonly been used in speaker adaptation. In the computation of the transformation matrix usually only one iteration of the Expectation-Maximisation (EM) algorithm is used, but there is not a complete study about results with a different number of iterations. We analyze how the number of iterations affects to the adaptation. The obtained results lead us to suggest a new method to accelerate the convergence of adaptation. Additionally, we propose a way to verify the contribution of the different adaptation matrices obtained in the EM process. We present experiments with the Wall Street Journal corpus whose aim is to determine the best option for the MLLR technique with respect to the number of EM iterations and the quality of the new convergence criterion.

Index Terms: speaker adaptation, speech recognition

1. Introduction

We can find Automatic Speech Recognition (ASR) systems all around the world. Current state-of-the-art ASR systems are based on Hidden Markov Models (HMM) to model the acoustic knowledge and n-grams to model the syntactical knowledge [1]. A robust ASR system needs to perform well in different environments and with different speakers. However, many speech recognition systems are for personal use, as only one speaker does usually use them (e.g., in a mobile phone or in a car). Consequently, it is interesting to guarantee an optimal performance for a particular speaker of an ASR system initially designed for multiple speakers. For this reason, speaker adaptation has become an essential part of a state-of-the-art ASR system.

An initial speaker-independent system can be adapted by using the Maximum Likelihood Linear Regression (MLLR) technique [2]. MLLR computes a set of transformations that reduces the mismatch between the initial model set and the adaptation data. These transformations are obtained by solving an optimisation problem using the Expectation-Maximisation (EM) technique [3]. The EM algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.

In the MLLR technique every iteration of the EM process provides a complete set of transformations that can be used to adapt the model. However, the MLLR technique has commonly been used with only one iteration of adaptation [2]. Some works study the performance of the adaptation with only the first iterations of the EM algorithm [4, 5, 6].

In this work we study the performance of the adaptation with respect to the number of iterations of the EM algorithm, from the first iteration to the convergence of the EM process. Since in each iteration the transformation matrix is closer to the identity matrix (because models are closer to adaptation data), we define a new stop criterion based on the distance of the transformation matrix to the identity matrix. Determining the contributions of the different matrices computed in the EM process can be used to verify whether the main contribution is given by the matrix of the first iteration. To determine these contributions we present a way to compute a general transformation matrix. We present experimental results on 8 speakers from the Wall Street Journal corpus to study this influence.

2. The MLLR adaptation technique

The aim of speaker adaptation techniques is to obtain a speaker-dependent recognition system by using a combination of general speech knowledge from well-trained HMM and speaker-specific information from a new speaker's data.

MLLR is a technique to adapt a set of speaker-independent acoustic models to a speaker by using small amounts of adaptation material. The MLLR approach requires an initial independent continuous density HMM system. MLLR adapts the acoustic models and updates the model mean parameters to maximise the likelihood of the adaptation data by using a transformation matrix, which is estimated from the adaptation data.

The theory is based on the concept of regression classes. A regression class is a set of mixture components that share the same transformation matrix $\vec{W}$. This matrix is applied to the extended mean vector of all the mixtures pertaining to the regression class to obtain an adapted mean vector. Given a state q in a HMM, for the ith gaussian of its output distribution, we denote its mean vector as $\vec{\mu}_{qi}$. The adapted mean vector $\widehat{\vec{\mu}}_{qi}$ is obtained by:

$$\widehat{\vec{\mu}}_{qi} = \vec{W} \cdot \vec{\xi}_{qi}$$

where $\widehat{\vec{\mu}}_{qi}$ is the adapted mean and $\vec{\xi}_{qi}$ is the extended mean vector defined as $\vec{\xi}_{qi} = [w, \mu^1_{qi}, \ldots, \mu^n_{qi}]' = [w : \vec{\mu}_{qi}]$, where n is the number of features, $\vec{\mu}_{qi}$ is the original mean vector and w is an offset term.

If we have a set of adaptation data denoted by the sequence of acoustic feature vectors $\vec{X} = \vec{x}_1\vec{x}_2...\vec{x}_T, \vec{x}_t \in \mathbb{R}^D, t = 1, ..., T$, we can estimate the transformation matrix $\widehat{\vec{W}}$ using the maximum likelihood approach as:

$$\widehat{\vec{W}} = \max_{\vec{W}} p_{\widehat{\vec{\lambda}}}(\vec{X})$$

where $\widehat{\vec{\lambda}}$ defines the parameters of the adapted model.

The problem is solved by using the EM algorithm [3]. EM is an iterative method which alternates between performing:

1. An expectation (E) step: which computes an expectation of the log likelihood with respect to the current estimate of the distribution for the latent variables.
2. A maximization (M) step: which computes the parameters which maximise the expected log likelihood found on the E step. These parameters are used to determine the distribution of the latent variables in the next E step.

The main idea is to define an auxiliary function $Q(\vec{\lambda}, \widehat{\vec{\lambda}})$ in the step E as: $Q(\vec{\lambda}, \widehat{\vec{\lambda}}) = \sum_{\vec{\theta} \in \vec{\Theta}} p_{\vec{\lambda}}(\vec{X}, \vec{\theta}) \cdot \log(p_{\widehat{\vec{\lambda}}}(\vec{X}, \vec{\theta}))$ where $\vec{\theta}$ is a state sequence and $\vec{\Theta}$ is the set of all possible state sequences with length T. The auxiliary function depends on both the initial model $\vec{\lambda}$ and the adapted model $\widehat{\vec{\lambda}}$. Using this definition it can be shown that by successively defining a new model $\widehat{\vec{\lambda}}$ which maximises Q (in the step M), the auxiliary function has the property that the value of $p_{\widehat{\vec{\lambda}}}(\vec{X})$ will not decrease, which was the original objective. The auxiliary function is maximised in the standard way by differentiating and equating to zero. The solution of this equation is the set of transformation matrices $\vec{W}$, that are applied on $\vec{\lambda}$ to obtain $\widehat{\vec{\lambda}}$. Therefore, it is necessary to compute a transformation matrix for every iteration of the EM algorithm until the algorithm converges. Our convergence criterion is based on the difference between the values of Q in the current and the previous iteration. We assume convergence when this difference is 0 (for the used numeric precision).

To compute the transformation matrix (M step), we can use several variants. Details on the estimation of these variants can be consulted in [2]. In our case, we suppose different covariances for each distribution and full adaptation matrices.

3. MLLR issues

One important problem in MLLR is the large number of iterations required for the EM convergence with a full transformation matrix. We know that in each iteration the transformation matrix is closer to the identity matrix because models are closer to adaptation data. If a transformation matrix is not closer to the identity matrix that the matrix of the previous iteration it is possible that models are overfitted. With this idea, we propose a new method to stop the EM estimation: in each iteration, we calculate the distance between the transformation matrix and the identity matrix using the Euclidean Distance Matrix [7]:

$$\delta(\vec{W}, \vec{I}) = ||\vec{W} - \vec{I}||$$

When this distance is higher than the distance obtained with the matrix of the previous iteration we stop the adaptation. Since the adaptation of the models depends directly on the transformation matrix, this seems a good criterion to evaluate when is worth applying the transformation matrix. Therefore, we use this idea as a new method to stop the estimation. With this option, we reduce the computing time to obtain good adapted models.

Since we want to study the performance of the adaptation with respect to the number of iterations of the EM algorithm, we need to determine the contribution of all adaptation matrices obtained in the EM process. We can define that the transformations define a path in the representation space between the initial and final models, and this path covers a certain distance. Therefore, to determine the contribution of each EM step, the idea is to determine how much distance is covered in each iteration. To determine the contribution of a sequence of EM steps we present a way to compute a general transformation matrix that reflects the whole effect of the EM steps (i.e., we calculate a general matrix that applied to the initial models allows to obtain the adapted model in any iteration).

The calculation of a general matrix was not defined previously, as far as we know. This matrix can not be calculated as the product of the different matrices obtained in the EM process (because dimensions do not match). We have obtained a way to compute a general matrix to obtain the adapted models from the initial models. The method solves a linear equation system where the unknown variables are the matrix coefficients. These coefficients are calculated using only n means (where n is the number of rows of $\vec{W}$) to be transformed from the initial model to the adapted model. For example, for $n = 2$, if we have the original means $\vec{\varepsilon_1} = [w\ \varepsilon_{11}\ \varepsilon_{12}]$, $\vec{\varepsilon_2} = [w\ \varepsilon_{21}\ \varepsilon_{22}]$ and the corresponding adapted means $\vec{\mu_1} = [\mu_{11}\ \mu_{12}]$, $\vec{\mu_2} = [\mu_{21}\ \mu_{22}]$, the coefficients of $\vec{W} = [w_{10}\ w_{11}\ w_{12}; w_{20}\ w_{21}\ w_{22}]$ are obtained by solving:

$$\begin{bmatrix} w & \varepsilon_{11} & \varepsilon_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & w & \varepsilon_{11} & \varepsilon_{12} \\ w & \varepsilon_{21} & \varepsilon_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & w & \varepsilon_{21} & \varepsilon_{22} \end{bmatrix} \begin{bmatrix} w_{10} \\ w_{11} \\ w_{12} \\ w_{20} \\ w_{21} \\ w_{22} \end{bmatrix} = \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix}$$

where w_{ij} are the unknown variables.

Using the definition of the general matrix, we compute the distance covered by the whole transformation process as the distance between the general matrix of the models obtained in the convergence ($\vec{W_c}$) and the identity matrix. The contribution of each step is calculated as the projection on this path, using the distance between the general matrix of step i ($\vec{W_i}$) and $\vec{W_c}$ and $\vec{I}$ to define the projection.

Furthermore, with the use of the general matrix it is not necessary to keep the models of each iteration (since these models can be obtained by applying the general transformation matrix on the original models), with the consequent space saving.

4. Corpus

The experiments were performed using the Wall Street Journal (WSJ) database [8]. The ARPA WSJ corpus consists of samples of read texts drawn from WSJ publications recorded under high-quality conditions. We have used the Nov'92 (WSJ0) and Nov'93 (WSJ1) training data.

The initial HMMs have been trained with HTK [9] using just the WSJ0 training database composed of 84 speakers with a duration of 15 hours. The HMMs are word-internal triphones and gender independent. They are composed of $2.3k$ tied-states and the topology is left-to-right with loops. The number of gaussians per state is 16. We have also trained silence and inter-word silence models.

For adaptation experiments, we have used a set of eight speakers selected from the WSJ1 (Nov'93) training material [1]. We have only one regression class (a global transformation matrix). There are about 150 utterances available for each speaker. 50 sentences were used for adaptation and the remaining were used for testing. As language model, we used the standard $20k$ trigram grammar.

5. Experiments and Results

To analyze the results, we used the Word Error Rate (WER) as the evaluation measure. This measure computes the edit dis-

[1] Notice that this subcorpus is not the usual WSJ benchmark and baseline results might not be comparable with other works.

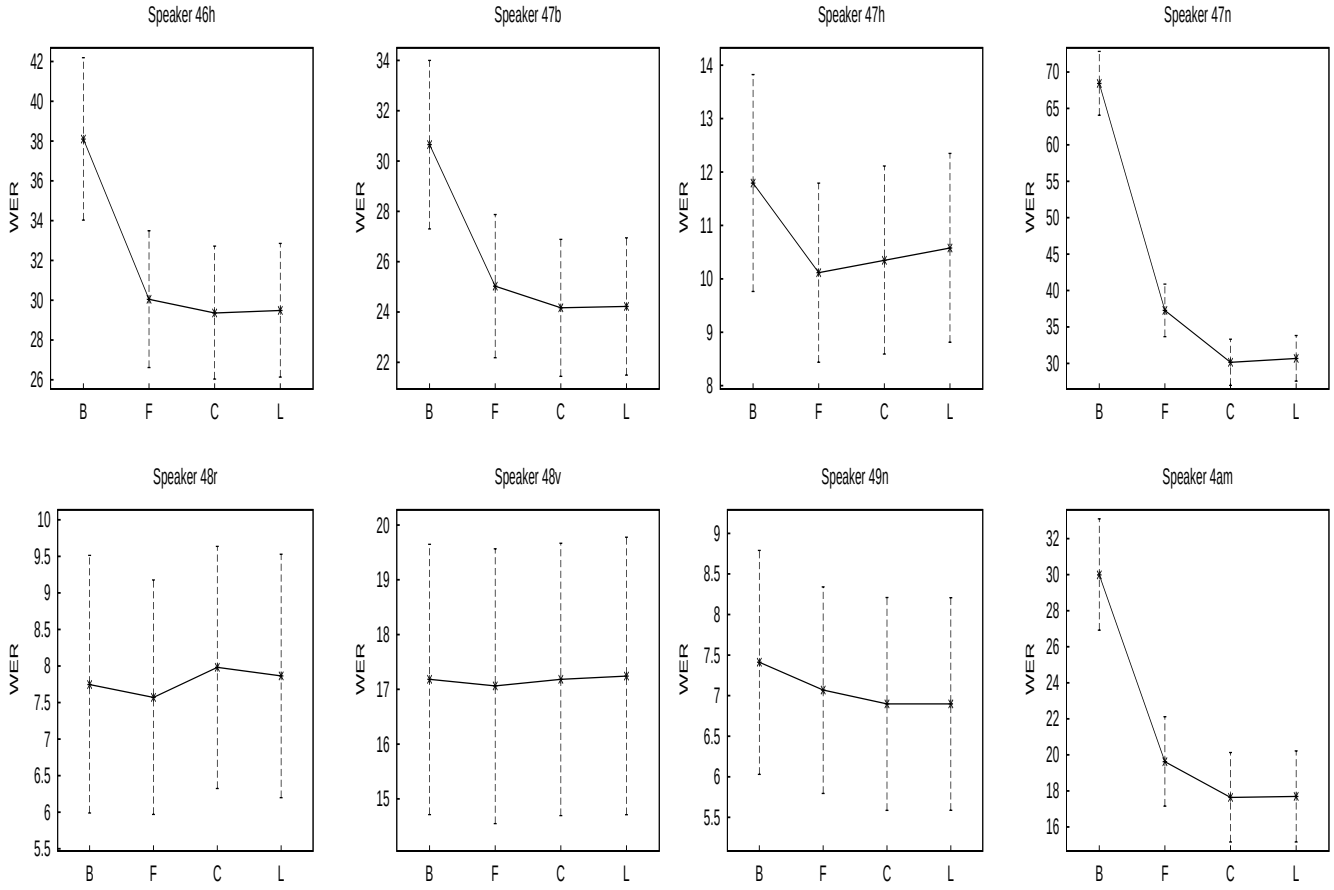

Figure 1: Comparison among full adaptation matrices with a different number of EM iterations for each speaker. B is baseline result. F is the result for the first iteration. C is the result for the matrix convergence. L is the result for the last iteration.

tance between a reference sentece and the recognized sentence.

We performed some experiments with the MLLR technique to determine its behavior with respect to the number of EM iterations. These experiments included a comparison between the recognition results obtained when applying the different transformations obtained in each EM iteration (from first iteration till convergence and the new convergence criterion). Recognition was carried out with the iATROS recogniser [10].

In Figure 1 we can see the results for each speaker with confidence intervals that show whether the differences among the results are statistically significant [11]. Every graphic shows four results:

- **Baseline**: It is the WER obtained when using the models without adaptation.
- **First iteration**: It is the WER obtained when using an adapted model with only one iteration of the EM algorithm.
- **Matrix convergence**: It is the WER obtained with an adapted model with the matrix convergence that we defined with the distance between the transformation matrix and the identity matrix.
- **Last iteration**: It is the WER obtained with a model adapted with the transformations obtained in the EM convergence.

According to these results, we can distinguish two groups of speakers:

- 46h, 47b, 47n, 4am: these speakers have a bad WER (above 30) when they use models without adaptation. These speakers have a better statistically significant WER when they use adapted models (47b has a statistically significant WER only for the matrix convergence or last iteration).
- 47h, 48r, 48v, 49n: these speakers have a better WER (below 20) when they use models without adaptation. In these speakers it is not necessary to adapt the original models because the adaptation does not improve the recognition results (the differences among the results are not statistically significant). With these speakers we can not draw conclusions.

In Figure 2 we show the results that we obtain if we calculate the mean of all speakers. The results are better than baseline because confidence intervals show that the differences among the results are statistically significant. Although the results when using EM till convergence (with matrix or EM convergence criteria) are not significantly improved with respect to the results of the first iteration, absolute results are slightly improved and show that it could be convenient to use more than one EM iteration. The matrix convergence is a good option because the number of iterations is quite lower than in the EM convergence (for example, the speaker 46h has 606 iterations with the EM convergence and 19 iterations with the matrix convergence) but results are very similar and overfitting is possibly avoided.

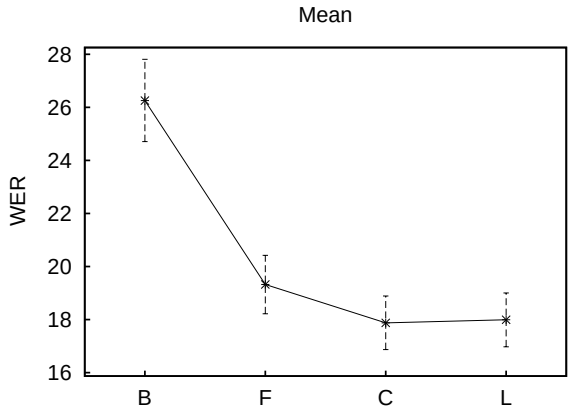

Figure 2: Mean WER for all the 8 speakers. B, F, C, and L have the same meaning than in Figure 1.

Figure 3: Distance and projections of each general adaptation matrix with respect to the identity matrix and the convergence adaptation matrix for speaker 46h.

With respect to the contributions of each EM step to the adaptation, in Figure 3 we show the calculated projections for the general matrices obtained in each EM step for speaker 46h. We can see that the first distance is the highest distance of all distances. Therefore, we think that the matrix of the first iteration is the matrix that produces a more important transformation in the data and it is the matrix with a greater contribution in the adaptation process. Since the distance decreases in every iteration, the contribution is lower and the first iterations are the most important in the adaptation process.

6. Conclusions and Future Work

The results show that MLLR speaker adaptation significantly improves the performance in speakers with bad performance with speaker independent acoustic models. The results show that there is no significant improvement between using only one EM iteration or more iterations in the recognition performance. We presented a new EM convergence criterion that obtains adapted models with similar performance to those obtained with the usual EM stop criterion and that are possibly not overfitted. Moreover, the number of iterations of EM is lower when using this new criterion. Therefore, we reduced the time of computation.

We studied the contribution of the adaptation matrices to confirm that the transformation matrix of the first iteration is that which produces the major contribution. We demonstrated it with an empirical method, using the recognition results and the distance between matrices. To perform this analisys we provided a method to calculate a general matrix that computes the adapted models from initial models. Furthermore, with the general transformation matrix we reduce the memory to apply adaptation, since only the initial models and a transformation matrix are needed to obtain the adapted models.

Future work is directed towards using more regression classes and automatic building of regression classes. Moreover, other convergence criteria can be defined and experiments can be performed with other corpora to confirm these conclusions. The use of MLLR on handwritten text recognition in another interesting work.

7. Acknowledgements

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV “Consolider Ingenio 2010” program (CSD2007-00018) and MITTRAL (TIN2009-14633-C03-01) projects. Also supported by the EC (FEDER) and the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project, by the Generalitat Valenciana under grants Prometeo/2009/014, GV/2010/067, ACOMP/2010/051 and by VIDI-UPV under PAID06-20070315 program.

8. References

[1] L. Rabiner and B. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice Hall PTR.

[2] C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9:171–185.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38.

[4] C. J. Leggetter. 1995. Improved Acoustic Modeling for HMMs using Linear Transformations. *PhD thesis, Cambridge University Engineering Department*.

[5] J. Kleban and Y. Gong 2000. HMM adaptation and microphone array processing for distant speech recognition. In *Proceedings of the 2000 IEEE international Conference on Acoustics, Speech and Signal Processing*, ICASSP. IEEE Computer Society, vol.3.

[6] T. Shinozaki, Y. Kubota and S. Furui. 2009. Unsupervisec cross-validation adaptation algorithms for improved adaptation performance. In *Proceedings of the 2009 IEEE international Conference on Acoustics, Speech and Signal Processing*, ICASSP. IEEE Computer Society, Washington, DC, 4377-4380.

[7] J. Dattorro. 2005. Convex Optimization and Euclidean Distance Geometry. *Meboo Publishing*.

[8] D. S. Pallett, J. G. Fiscus, W. M. Fisher, John S. Garofolo, B. A. Lund, and M. A. Przybocki. 1994. 1993 benchmark tests for the arpa spoken language program. In *HLT '94: Proceedings of the workshop on HLT*, pages 49–74, Morristown, NJ, USA.

[9] Young, S., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, July, 2004. *The HTK Book*. CUED, UK, v3.2 edition.

[10] M. Luján-Mares, V. Tamarit, V. Alabau, C. D. Martínez-Hinarejos, M. Pastor, A. Sanchis, and A. Toselli. 2008. iatros: A speech and handwritting recognition system. In *V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78, Spain.

[11] M. Bisani and H. Ney. 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *Proceedings of ICASSP'04*, volume 1, pages 409–412.