# Initialization of fMLLR with Sufficient Statistics from Similar Speakers*

Zbyněk Zajíc, Lukáš Machlica, and Luděk Müller

University of West Bohemia in Pilsen,
Faculty of Applied Sciences, Department of Cybernetics,
Univerzitní 22, 306 14 Pilsen
{zzajic,machlica,muller}@kky.zcu.cz

**Abstract.** One of the most utilized adaptation techniques is the feature Maximum Likelihood Linear Regression (fMLLR). In comparison with other adaptation methods the number of free parameters to be estimated significantly decreases. Thus, the method is well suited for situations with small amount of adaptation data. However, fMLLR still fails in situations with extremely small data sets. Such situations can be solved through proper initialization of fMLLR estimation adding some a-priori information. In this paper a novel approach is proposed solving the problem of fMLLR initialization involving statistics from speakers acoustically close to the speaker to be adapted. Proposed initialization suitably substitutes missing adaptation data with similar data from a training database, fMLLR estimation becomes well-conditioned, and the accuracy of the recognition system increases even in situations with extremely small data sets.

**Keywords:** fMLLR, adaptation, sufficient statistics, speech recognition, robustness, initialization.

## 1 Introduction

Nowadays, in Automatic Speech Recognition (ASR) systems speaker adaptation of an acoustics model represents a standard approach how to improve the accuracy of an ASR system. The most widely used method is feature Maximum Likelihood Linear Regression (fMLLR), which transforms acoustic features for better fit to a Speaker Independent (SI) model.

fMLLR tries to find a linear transformation ($N \times N$ matrix, where $N$ is the features dimension) of an acoustic space, which maximizes probability of test data given a SI model. In the case where small amount of adaptation data is available (especially in on-line recognition) the number of free parameters ($N \times N$) is too high to be properly estimated. Transformation matrix becomes ill-conditioned, and can lead to poor recognition. Some solutions of the problem were already proposed, e.g. lower the number of free parameters using diagonal matrices [1], eigenspace approach [2], or initialization of the estimation [3]. Initialization methods suppress the influence of adaptation data

* This research was supported by the Ministry of Education of the Czech Republic project No. MŠMT LC536, by the Grant Agency of the Czech Republic project No. GAČR 102/08/0707, and the grant of The University of West Bohemia project No. SGS-2010-054.

I. Habernal and V. Matoušek (Eds.): TSD 2011, LNAI 6836, pp. 187–194, 2011.

for the benefit of initialization data. Usually a compromise has to be made between safety and accuracy of the adaptation. Next problem to solve is how to choose suitable initialization data.

In this paper a similar approach to [4] is used, where prearranged data statistics from similar speakers are utilized to perform an additional EM (Expectation-Maximization) iteration of the SI model according to given statistics, see Section 4. We use the same statistics, but in order to initialize the fMLLR estimation, what proves to be efficient mainly in cases with extremely small amount of adaptation data. The selection of closest (most similar) speakers to the test speaker is crucial and is described in more detail in Section 5. Results of the proposed method compared with standard (basic) fMLLR and unadapted SI system for different sizes of adaptation data can be found in Section 6.

## 2 Adaptation

The adaptation adjusts the SI model so that the probability of the adaptation data would be maximized. The most widely used methods for adaptation are Maximum A-posteriori Probability (MAP) technique and Linear Transformations (LTs). Adaptation techniques do not access the data directly, but only through accumulated statistics, which is the first step preceding the adaptation process. Instead of storing a huge amount of data to estimate the adaptation formulas, adaptation methods need only following statistics:

$$\gamma_{jm}(t) = \frac{\omega_{jm} p(\boldsymbol{o}_t|jm)}{\sum_{m=1}^{M} \omega_{jm} p(\boldsymbol{o}_t|jm)} \tag{1}$$

denoting the $m-th$ mixture's posterior of the $j-th$ state of the HMM,

$$c_{jm} = \sum_{t=1}^{T} \gamma_{jm}(t) \tag{2}$$

representing the soft count of mixture $m$,

$$\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}) = \textstyle\sum_{t=1}^{T} \gamma_{jm}(t)\boldsymbol{o}_t\,, \quad \boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) = \textstyle\sum_{t=1}^{T} \gamma_{jm}(t)\boldsymbol{o}_t\boldsymbol{o}_t^{\mathrm{T}} \tag{3}$$

denoting the sum of the first and the second moment of features aligned to mixture $m$ in the $j$-th state of the HMM.

## 3 Feature Maximum Likelihood Linear Regression (fMLLR)

fMLLR technique belongs to the category of Linear Transformations (LTs), another LT based method is Maximum Likelihood Linear Regression (MLLR). These methods try to find a linear transformation in order to match adaptation data with an acoustics model. Contrary to MAP, LTs can adapt more model components at once using the same transformation (e.g. only one matrix for all the model means), thus they require lower amount of adaptation data since number of free parameters to be estimated is low. Similar model components are clustered into clusters $K_n, n = 1, \ldots, N$ in order to

lower the number of adapted parameters [7]. Advantage of fMLLR is that it transforms directly acoustics features instead of an acoustics model (this is the case for MLLR), what is less time-consuming. fMLLR transforms features $\boldsymbol{o}_t$ according to

$$\bar{\boldsymbol{o}}_t = \boldsymbol{A}_{(n)}\boldsymbol{o}_t + \boldsymbol{b}_{(n)} = \boldsymbol{W}_{(n)}\boldsymbol{\xi}(t)\,, \tag{4}$$

where

$$\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}], \tag{5}$$

$\boldsymbol{W}_{(n)}$ represents the transformation matrix corresponding to the $n-th$ cluster $K_n$ and $\boldsymbol{\xi}(t) = [\boldsymbol{o}_t^{\mathrm{T}}, 1]^{\mathrm{T}}$ stands for the extended feature vector.

The estimation formulas of rows of $\boldsymbol{W}_{(n)}$ are given as

$$\boldsymbol{w}_{(n)i} = \boldsymbol{G}_{(n)i}^{-1}\left(\frac{\boldsymbol{v}_{(n)i}}{\alpha_{(n)}} + \boldsymbol{k}_{(n)i}\right)\,, \tag{6}$$

where $\boldsymbol{v}_{(n)i}$ is the $i$-th row vector of cofactors of matrix $\boldsymbol{A}_{(n)}$, $\alpha_{(n)}$ can be found as a solution of a quadratic function defined in [8],

$$\boldsymbol{k}_{(n)i} = \sum_{m\in K_n} \frac{\mu_{mi}\boldsymbol{\varepsilon}_m(\boldsymbol{\xi})}{\sigma_{mi}^2}\,, \quad \boldsymbol{G}_{(n)i} = \sum_{m\in K_n} \frac{\boldsymbol{\varepsilon}_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}})}{\sigma_{mi}^2}\,, \tag{7}$$

where $\boldsymbol{G}_{(n)i}$, $\boldsymbol{k}_{(n)i}$ are accumulation matrices of statistics (3) of all mixtures $m$ contained in a given cluster $K_n$, and

$$\boldsymbol{\varepsilon}_m(\boldsymbol{\xi}) = \left[\boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}), c_m\right]^{\mathrm{T}}\,, \quad \boldsymbol{\varepsilon}_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}) = \begin{bmatrix} \boldsymbol{\varepsilon}_m(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) & \boldsymbol{\varepsilon}_m(\boldsymbol{o}) \\ \boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}) & c_m \end{bmatrix}\,. \tag{8}$$

Equation (6) is a solution of the minimization problem with auxiliary function given in [8]. Matrices $\boldsymbol{A}_{(n)}$ and $\boldsymbol{b}_{(n)}$ are estimated iteratively, thus they have to be suitably initialized (for enough data e.g. randomly, otherwise see Section 5).

In order to accumulate (7) each frame has to be assigned to a HMM state in the first place. In the case of unsupervised fMLLR, the assignment process (more precisely the recognition) has to be done utilizing the not adapted SI system. Since the recognition may contain errors it is suitable to assign a Certainty Factor (CF $\in \langle 0, 1\rangle$) to each of the recognized phones/words/sentences/etc. We work on the word level, CFs for any particular word sequence are extracted from the lattice and may be computed as in [9]. We use only the best path in the lattice. Only the data which transcriptions have high CF (greater than an empirically specified threshold) are used for the adaptation. Still some problems may occur. Even if the CF of a word is high, the boundaries of the word can be inaccurate, because of low values of CF of neighborhood words. Hence, it is useful to take into account the left and right context of each word in the sense of CF. We are seeking for a sequence of three words, where each of them has a CF higher than the threshold and for adaptation we consider only the middle one.

## 4 Sufficient Statistics of Closest Speakers

This method was proposed in [4]. It consists in selecting a subset of speakers who are close in the acoustic space to a given test speaker $t$ whose speech is going to be

recognized. Model of the $t$ - th speaker is then estimated according to the already stored HMM data statistics of selected speakers. Compared to the Cluster Adaptive Training (CAT) [5], [6] this method can result in more suitable clusters, because the clusters are determined according to the data of an actual test speaker, hence on-line. This method consists of three steps – accumulation of statistics, cohort selection and estimation of a new model (see Fig.1).

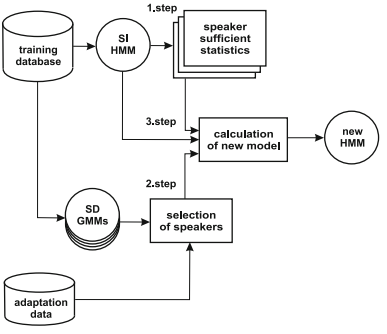


**Fig. 1.** Three steps in the process of a SI model reestimation based on sufficient statistics of closest speakers. SI stands for Speaker Independent model, SD for Speaker Dependent model.

### 4.1 Accumulation of Sufficient Statistics

In the first step, given a SI model sufficient statistics (1)-(3) are accumulated for each speaker in the training database using all his speech data. Hence a set of sufficient statistics is acquired. This step can be done off-line, before the adaptation itself.

### 4.2 Selecting a Cohort of Speakers

In the second step $N$-best speakers are selected from the training database according to their closeness to the test speaker $t$. The selection is performed on-line utilizing $t$ - th speaker's present data. For each of the training speakers a 64 mixture GMM is trained from all their available data. Then, $t$ - th speakers data are scored against each of the GMMs. Verification scores are sorted and $N$-best speakers with highest likelihoods are included into the cohort.

### 4.3 Estimating a New Model

The statistics from the speakers in the cohort are used to reestimate the SI model in order to better match the $t$-th speaker actual data. In [4] one EM iteration of the HMM is performed.

## 5 fMLLR Initialization through Sufficient Statistics

In this section we combine both approaches described above – fMLLR adaptation from Section 3 and statistics of closest speakers from Section 4. We do not train a new model,

instead we use the data statistics in order to initialize the fMLLR estimation process. The principle is similar to the steps defined in Section 4 except for a few differences.

### 5.1 Accumulation of Sufficient Statistics

First, for each speaker $s$ in the training set directly matrices $\boldsymbol{k}^s_{(n)i}$ and $\boldsymbol{G}^s_{(n)i}$ given in (7) are accumulated and stored.

### 5.2 Selecting a Cohort of Speakers

Second, we do not compute the likelihood of the whole data set (utterance/recording) of a test speaker $t$ at once. Instead we use a floating window with a fixed size and a fixed shift. All the frames in a window are scored against all the GMMs trained for each of the speakers in the training set and then the window shifts to a new position. For each window position the best scoring GMM is found and the corresponding statistics (matrices $\boldsymbol{k}^s_{(n)i}$ and $\boldsymbol{G}^s_{(n)i}$) are added to the cohort. Thus, the size of the cohort $N_t$ changes for each test speaker $t$, it is not fixed as in [4]. In order to train the GMMs we use MAP adaptation of an Universal Background Model (UBM) [10]. UBM also participates in the scoring of frames in a window, however if UBM scores best nothing is added into the cohort. It should serve to discard non-informative frames.

### 5.3 fMLLR Transform Estimation

Third, fMLLR transformation matrix (5) is computed using all the matrices from the cohort and statistics obtained from the $t$ - th speaker's available data, thus:

$$\boldsymbol{k}_{(n)i} = \sum_{s=1}^{N_t} \boldsymbol{k}^s_{(n)i} + \boldsymbol{k}^t_{(n)i} \ , \quad \boldsymbol{G}_{(n)i} = \sum_{s=1}^{N_t} \boldsymbol{G}^s_{(n)i} + \boldsymbol{G}^t_{(n)i} \ , \tag{9}$$

for each cluster $n$ and each row $i$ of resulting $\boldsymbol{W}_{(n)}$.

## 6 Experiments

### 6.1 SpeechDat-East (SD-E) Corpus

SpeechDat-East (see [12]) contains telephone speech in 5 languages Czech, Polish, Slovak, Hungarian, and Russian. We used only the Czech part of SD-E. In order to extract the features Mel-frequency cepstral coefficients (MFCC) were utilized, 11 dimensional feature vectors were extracted each 10 ms utilizing a 32 ms hamming window, Cepstral Mean Normalization (CMN) was applied, and $\Delta$, $\Delta^2$ coefficients were added.

A 3 state HMM based on triphones with 2105 states total and 8 GMM mixtures with diagonal covariances in each of the states was trained on 700 speakers with 50 sentences for each speaker (cca 5 sec. for sentence). Using the same data 256 mixture UBM was trained and subsequently all the GMMs of individual speakers were MAP adapted.

To test the systems performance different 200 speakers from SDE were used with 50 sentences for each speaker, however 12 sentences maximal were used for the adaptation. For the recognition a language model based on trigrams was considered [11]. The vocabulary consisted of 7000 words.

### 6.2 Adaptation Setup

In our experiments we utilized unsupervised fMLLR adaptation. At first, the SI model was used to get the word transcription with assigned CFs (as described in Section 3) of given sentences. At the same time a cohort containing statistics was determined as described in Section 5. The window size was set to 30 frames with 10 frames shift resulting in cca 20 speakers per cohort. Note that in a case when several sentences of a test speaker were available we used only the first sentence to determine the cohort. This makes the adaptation process significantly faster preserving sufficient amount of statistics in the cohort (assuming 5 sec. sentences and floating window of size 30 frames with 10 frames shift) – good compromise between good results and small time consumption (important in on-line recognition). At the end, statistics from the cohort along with statistics of given sentences of a test speaker corresponding to words with adequate CFs (see Section 3) were used to estimate the fMLLR transformation matrices. CF threshold was set to 0.99, usable length of one adaptation sentence is then shorter than proclaimed 5 sec. (cca 3 sec.). At the end, the given sentences were once again recognized with adapted model.

Clustering of model components was performed via a regression tree. The threshold for occupation of nodes in the regression tree was set to 1000. In basic (standard) fMLLR the number of transformation matrices depends on the number of adaptation sentences. In the case of fMLLR initialized with sufficient statistics of $N_t$-best speakers ($N_t$best-fMLLR) one can determine the number of matrices in advance since there is always enough adaptation data guaranteed by the cohort. For the $N_t$best-fMLLR 32 transformation matrices for each speaker were created. Only one iteration of fMLLR was carried out.

### 6.3 Results

Results of experiments on basic fMLLR and $N_t$best-fMLLR in dependence on varying number of adaptation sentences can be found in Table 1 and in Figure 2. Results show poor performance of basic fMLLR on small sets of adaptation data, where the estimation of transformation parameters is ill-conditioned (for 5 and less adaptation

**Table 1.** Accuracy (Acc)[%] of the unadapted SI system (baseline), adapted system with basic fMLLR and fMLLR initialized by sufficient statistics from $N_t$best speakers ($N_t$best-fMLLR) in dependence on the number of adaptation sentences. Real system combines basic $N_t$best-fMLLR and fMLLR, $N_t$best-fMLLR is replaced by fMLLR after cca 20-30 sec. of adaptation speech.

| Number of sentences | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|---|---|---|
| unadapted SI system | 62.69 | 62.69 | 62.69 | 62.69 | 62.69 | 62.69 | 62.69 | 62.69 | 62.69 |
| basic fMLLR | 12,31 | 50.49 | 59.73 | 60.62 | 62.03 | 62.82 | 63.71 | 64.19 | 64.25 |
| $N_t$best-fMLLR | 62.77 | 63.37 | 63.31 | 62.96 | 63.13 | 63.22 | 63.23 | 63.54 | 63.65 |
| real system | 62.77 | 63.37 | 63.31 | 62.96 | 63.13 | 63.22 | 63.71 | 64.19 | 64.25 |

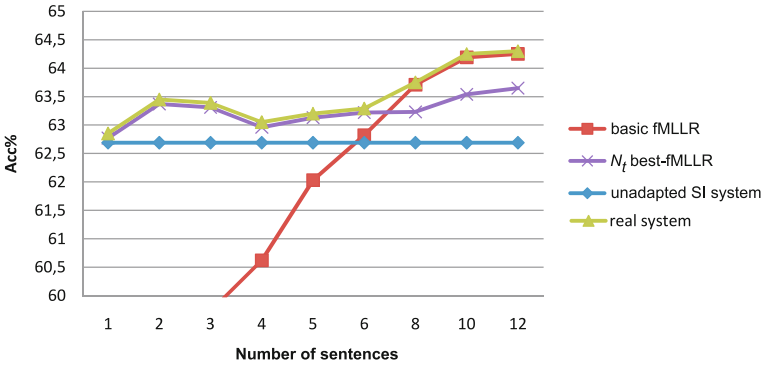


**Fig. 2.** Accuracy (Acc)[%] of systems in dependence on the number of adaptation sentences, see also Table 1

sentences). This is not true for $N_t$best-fMLLR, which actually outperforms the baseline even for small amounts of data. Hence, statistics from closest speakers in the cohort ensure proper initialization and can be thought as a good approximation of statistics of utterances of the test speaker. When the number of sentences (speech data) increases, basic fMLLR becomes better than the $N_t$best-fMLLR.

This is caused by the fact that the overall statistics composed of the test speaker's statistics and statistics of cohort speakers make the final transform more speaker independent than the basic fMLLR. The solution is simple, a threshold has to be determined where the amount of data is sufficient in order to discard the cohort statistics from the estimation process (real system in Figure 2).

## 7 Conclusion

In this paper we proposed an initialization method based on sufficient statistics from $N_t$best speakers ($N_t$best-fMLLR) in order to prevent the poor performance of basic fMLLR in cases of small adaptation data sets. Proposed method uses off-line accumulated statistics form closest speakers in the acoustic space to stabilize the estimation of transformation matrices. Every initialization has to cope with a compromise between better accuracy and well-conditioned adaptation since initialization data hold back the impact of adaptation data. Results show that assuming only few adaptation data $N_t$best-fMLLR outperforms basic fMLLR, because missing data are replaced by $N_t$ best speaker's sufficient statistics. For two adaptation sentences $N_t$best-fMLLR improves the accuracy by 0.6% absolutly compared to unadapted SI system, while fMLLR worstens the performance. Assuming more than 6 adaptation sentences fMLLR gains the lead since the $N_t$ best statistics make the adaptation more speaker independent. Hence, it is useful to lower the number of initialization statistics according to the quantity of adaptation data.

## References

1. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12, 75–98 (1997)
2. Chen, K., Liau, W., Wang, H., Lee, L.: Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. In: International Conference on Spoken Language Processing, Beijing, China, pp. 742–745 (2000)
3. Li, Y., et al.: Incremental on-line feature space MLLR adaptation for telephony speech recognition. In: International Conference on Spoken Language Processing, Denver (2002)
4. Yoshizawa, S., Baba, A., Matsunami, K., Mera, Y., Yamada, M., Shikano, K.: Unsupervised speaker adaptation based on sufficient HMM statistics of selected speakers. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 341–344 (2001)
5. Gales, M.J.F.: Cluster adaptive training of hidden Markov models. IEEE Transactions on Speech and Audio Processing, 417–428 (2000)
6. Vaněk, J., Psutka, J., Zelinka, J., Trmal, J.: Training of speaker-clustered acoustic models for use in real-time recognizers. In: Sigmap 2009, Milan, pp. 131–135 (2009)
7. Gales, M.J.F.: The generation and use of regression class trees for MLLR adaptation. Cambridge University Engineering Department, Cambridge (1996)
8. Povey, D., Saon, G.: Feature and model space speaker adaptation with full covariance Gaussians, Interspeech, paper 2050-Tue2BuP.14 (2006)
9. Uebel, L.F., Woodland, P.C.: Improvements in linear transform based speaker adaptation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 49–52 (2001)
10. Reynolds, D. A., Quatieri, T. F., Dunn, R. D.:Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, 19–41 (2000)
11. Pražák, A., Psutka, J., Hoidekr, J., et al.: Automatic online subtitling of the Czech parliament meetings. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 501–508. Springer, Heidelberg (2006)
12. Pollak, P., et al.: SpeechDat(E) - Eastern European Telephone Speech Databases. In: XLDB - Very Large Telephone Speech Databases (ELRA), Paris (2000)