Refinement Approach for Adaptation Based on Combination of MAP and fMLLR

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. This paper deals with a combination of basic adaptation techniques of Hidden Markov Model used in the speech recognition. The adaptation methods approach the data only through their statistics, which have to be accumulated before the adaptation process. When performing two adaptations subsequently, the data statistics have to be accumulated twice in each of the adaptation passes. However, when the adaptation methods are chosen with care, the data statistics may be accumulated only once, as proposed in this paper. This significantly reduces the time consumption and avoids the need to store all the adaptation data. Combination of Maximum A-Posteriori Probability and feature Maximum Likelihood Linear Regression adaptation is considered. Motivation for such an approach could be the on-line adaptation, where the time consumption is of big importance.

1 Introduction

Nowadays, systems of speech recognition are based on Hidden Markov Models (HMMs) with output probabilities described mainly by Gaussian Mixture Models (GMMs) [1]. To recognize the speech from a recording one could train a Speaker Dependent (SD) model for each of the speakers present in the recording. However, this is in praxis often intractable because of the need of a large database of utterances coming from one speaker. Instead, so called Speaker Independent (SI) model is trained from large amount of data collected from many speakers, and subsequently, the SI model is adapted to better capture the voice of the talking person. Thus, a SD model is acquired.

More precisely, the adaptation adjusts the SI model so that the probability of the adaptation data would be maximized. Well known adaptation methods are Maximum A-posteriori Probability (MAP) technique (see Section 2.2) and Linear Transformations based on Maximum Likelihood (LTML) (see Section 2.3). All these methods address the data undirectly through their statistics defined in Section 2.1. Because MAP and LTML adaptation work in different way, it would be suitable to combine them in order to gain a superior performance. One of the possibilities would be to accumulate data statistics utilizing the SI model and the adaptation data, run MAP adaptation, accumulate new data statistics based on the MAP adapted model and the same data, and perform one of the LTML adaptations (or vice versa – perform LTML after MAP). Obviously, the main disadvantage consists in the need to store all the adaptation data

V. Matoušek and P. Mautner (Eds.): TSD 2009, LNAI 5729, pp. 274–281, 2009.

and run the system twice. This causes increased time consumption dependent on the amount of processed data. Therefore, in Section 2.4 we proposed an efficient method that avoids the need to accumulate the data statistics twice. In this paper we have chosen out of LTML based adaptations preferably the feature transformations because of the implementation issues (see Section 2.4). In addition, the feature transformations are well suited for on-line adaptation, see [9]. Experimental results discussed in Section 3.2 prove the suitability of the proposed method.

2 Adaptation Techniques

The difference between the adaptation and ordinary training methods stands in the prior knowledge about the distribution of model parameters, usually derived from the SI model [2]. The adaptation adjusts the model in order to maximize the probability of adaptation data. Hence, the new, adapted parameters can be chosen as

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} p(\boldsymbol{O}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}), \tag{1}$$

where $p(\boldsymbol{\lambda})$ stands for the prior information about the distribution of the vector $\boldsymbol{\lambda}$ containing model parameters, $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$ is the sequence of T feature vectors related to one speaker, $\boldsymbol{\lambda}^*$ is the best estimation of parameters of the SD model. We will focus on HMMs with output probabilities of states represented by GMMs. GMM of the $j-th$ state is characterized by a set $\boldsymbol{\lambda}_j = \{\omega_{jm}, \boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}\}_{m=1}^{M_j}$, where M_j is the number of mixtures, ω_{jm}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{C}_{jm}$ are weight, mean and variance of the $m-th$ mixture, respectively.

The most know adaptation methods are Maximum A-posteriori Probability (MAP) [4] and Linear Transformations based on the Maximum Likelihood (LTML) [6], the description of these methods will be given in following sections.

2.1 Statistics of Adaptation Data

As will be shown soon, the adaptation techniques do not access the data directly, but only through some statistics, which are accumulated beforehand. Let us define these statistics:

$$\gamma_{jm}(t) = \frac{\omega_{jm} p(\boldsymbol{o}(t)|jm)}{\sum_{m=1}^{M} \omega_{jm} p(\boldsymbol{o}(t)|jm)} \tag{2}$$

stands for the $m-th$ mixtures' posterior of the $j-th$ state of the HMM,

$$c_{jm} = \sum_{t=1}^{T} \gamma_{jm}(t) \tag{3}$$

is the soft count of mixture m,

$$\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}) = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \boldsymbol{o}(t)}{\sum_{t=1}^{T} \gamma_{jm}(t)}, \quad \boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) = \frac{\sum_{t=1}^{T} \gamma_{jm}(t) \boldsymbol{o}(t) \boldsymbol{o}(t)^{\mathrm{T}}}{\sum_{t=1}^{T} \gamma_{jm}(t)} \tag{4}$$

represent the first and the second moment of features which align to mixture m in the j-th state of the HMM. Note that $\boldsymbol{\sigma}_{jm}^2 = \mathrm{diag}(\boldsymbol{C}_{jm})$ is the diagonal of the covariance matrix $\boldsymbol{C}_{jm}$.

2.2 Maximum A-posteriori Probability (MAP) Adaptation

MAP is based on the Bayes method for estimation of the acoustic model parameters, with the unit loss function [3]. MAP adapts each of the parameters separately, therefore it is necessary to have for all the parameters enough adaptation data. The result of adaptation is negligible for small amount of data. The parameters are adapted according to formulas

$$\bar{\omega}_{jm} = [\alpha_{jm} c_{jm}/T + (1 - \alpha_{jm})\omega_{jm}]\chi \,, \tag{5}$$

$$\bar{\boldsymbol{\mu}}_{jm} = \alpha_{jm}\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}) + (1 - \alpha_{jm})\boldsymbol{\mu}_{jm} \,, \tag{6}$$

$$\bar{\boldsymbol{C}}_{jm} = \alpha_{jm}\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) + (1 - \alpha_{jm})(\boldsymbol{\sigma}^2_{jm} + \boldsymbol{\mu}_{jm}\boldsymbol{\mu}^{\mathrm{T}}_{jm}) - \bar{\boldsymbol{\mu}}_{jm}\bar{\boldsymbol{\mu}}^{\mathrm{T}}_{jm} \,, \tag{7}$$

$$\alpha_{jm} = \frac{c_{jm}}{c_{jm} + \tau} \,, \tag{8}$$

where α_{jm} is the adaptation coefficient, which controls the balance between the old and new parameters using empirically determined parameter τ. The parameter τ determines how much the new data have to be "observed" in each mixture till the mixture parameters change (they shift in the direction of new parameters) [4]. χ is a normalization factor, which guarantees that all the new weights of the mixture for one state sum to unity.

2.3 Linear Transformations Based on Maximum Likelihood

These methods are based on linear transformations. The advantage over the MAP technique is that the number of available model parameters is reduced via clustering of similar model components [8]. The transformation is the same for all the parameters from the same cluster $K_n, n = 1, \ldots, N$. Hence, less amount of adaptation data is needed. The first of the methods introduced by Leggeter in [5] is known as Maximum Likelihood Linear Regression (MLLR) and was further investigated by Gales, who introduced feature MLLR (fMLLR). The main difference between these two approaches stands in the area of their interest. MLLR transforms means and covariances of the model, whereas fMLLR transforms directly the acoustic feature vectors. The MLLR method is out of our interest and the adaptation formulas can be found in [5].

Feature Maximum Likelihood Linear Regression (fMLLR)
The method is based on the minimization of the auxiliary function [6]:

$$\begin{aligned} Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = const - \frac{1}{2}\sum_{jm}\sum_{t}\gamma_{jm}(t)(const_{jm} + \log|\boldsymbol{C}_{jm}| + \\ + (\bar{\boldsymbol{o}}(t) - \boldsymbol{\mu}_{jm})^{\mathrm{T}}\boldsymbol{C}^{-1}_{jm}(\bar{\boldsymbol{o}}(t) - \boldsymbol{\mu}_{jm})) \,, \end{aligned} \tag{9}$$

where $\bar{\boldsymbol{o}}(t)$ represents the feature vector transformed according to the formula:

$$\bar{\boldsymbol{o}}_t = \boldsymbol{A}_{(n)}\boldsymbol{o}_t + \boldsymbol{b}_{(n)} = \boldsymbol{W}_{(n)}\boldsymbol{\xi}(t) \,, \tag{10}$$

where $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$ stands for the transformation matrix corresponding to the $n-th$ cluster K_n and $\boldsymbol{\xi}(t) = [\boldsymbol{o}_t^{\mathrm{T}}, 1]^{\mathrm{T}}$ represents the extended feature vector. The auxiliary function (9) can be rearanged into the form [7]

$$Q_{\boldsymbol{W}_{(n)}}(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = \log|\boldsymbol{A}_{(n)}| - \sum_{i=1}^{I} \boldsymbol{w}_{(n)i}^{\mathrm{T}} \boldsymbol{k}_i - 0.5 \boldsymbol{w}_{(n)i}^{\mathrm{T}} \boldsymbol{G}_{(n)i} \boldsymbol{w}_{(n)i}\,, \tag{11}$$

where

$$\boldsymbol{k}_{(n)i} = \textstyle\sum_{m \in K_n} \frac{c_m \mu_{mi} \boldsymbol{\varepsilon}_m(\boldsymbol{\xi})}{\sigma_{mi}^2}\,, \quad \boldsymbol{G}_{(n)i} = \textstyle\sum_{m \in K_n} \frac{c_m \boldsymbol{\varepsilon}_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}})}{\sigma_{mi}^2} \tag{12}$$

and

$$\boldsymbol{\varepsilon}_m(\boldsymbol{\xi}) = \left[\boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}), 1\right]^{\mathrm{T}}\,, \quad \boldsymbol{\varepsilon}_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}) = \begin{bmatrix} \boldsymbol{\varepsilon}_m(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) & \boldsymbol{\varepsilon}_m(\boldsymbol{o}) \\ \boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}) & 1 \end{bmatrix}\,. \tag{13}$$

To find the solution of equation (11) we have to express $\boldsymbol{A}_{(n)}$ in terms of $\boldsymbol{W}_{(n)}$, e.g. use the equivalency $\log|\boldsymbol{A}_{(n)}| = \log|\boldsymbol{w}_{(n)i}^{\mathrm{T}} \boldsymbol{v}_{(n)i}|$, where $\boldsymbol{v}_{(n)i}$ stands for transpose of the $i-th$ row of cofactors of the matrix $\boldsymbol{A}_{(n)}$ extended with a zero in the last dimension. After the maximization of the auxiliary function (11) we receive

$$\frac{\partial Q(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}})}{\partial \boldsymbol{W}_{(n)}} = 0 \Rightarrow \boldsymbol{w}_{(n)i} = \boldsymbol{G}_{(n)i}^{-1} \left(\frac{\boldsymbol{v}_{(n)i}}{\alpha_{(n)}} + \boldsymbol{k}_{(n)i} \right)\,, \tag{14}$$

where $\alpha_{(n)} = \boldsymbol{w}_{(n)i}^{\mathrm{T}} \boldsymbol{v}_{(n)i}$ can be found as the solution of the quadratic function

$$\beta_{(n)} \alpha_{(n)}^2 - \alpha_{(n)}\, \boldsymbol{v}_{(n)i}^{\mathrm{T}} \boldsymbol{G}_{(n)i}^{-1} \boldsymbol{k}_{(n)i} - \boldsymbol{v}_{(n)i}^{\mathrm{T}} \boldsymbol{G}_{(n)i}^{-1} \boldsymbol{v}_{(n)i} = 0\,, \tag{15}$$

where

$$\beta_{(n)} = \sum_{m \in K_n} \sum_{t} \gamma_m(t)\,. \tag{16}$$

Two different solutions $\boldsymbol{w}_{(n)i}^{1,2}$ are obtained, because of the quadratic function (15). The one that maximizes the auxiliary function (11) is chosen. Note that an additional term appears in the log likelihood for fMLLR because of the feature transforms, hence:

$$\log \mathcal{L}\left(\boldsymbol{o}_t | \boldsymbol{\mu}_m, \boldsymbol{C}_m, \boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}\right) = \log \mathcal{N}\left(\boldsymbol{A}_{(n)} \boldsymbol{o}_t + \boldsymbol{b}_{(n)}; \boldsymbol{\mu}_m, \boldsymbol{C}_m\right) + 0.5 \log|\boldsymbol{A}_{(n)}|^2\,. \tag{17}$$

The estimation of $\boldsymbol{W}_{(n)}$ is an iterative procedure. Matrices $\boldsymbol{A}_{(n)}$ and $\boldsymbol{b}_{(n)}$ have to be correctly initialized first, e.g. $\boldsymbol{A}_{(n)}$ can be chosen as a diagonal matrix with ones on the diagonal and $\boldsymbol{b}_{(n)}$ can be initialized as a zero vector. The estimation ends when the change in parameters of transformation matrices is small enough (about 20 iterations are sufficient) [7].

2.4 Combination of MAP and fMLLR

As MAP and fMLLR work in different ways, it would be suitable to combine them. A simple method would be to run MAP and fMLLR subsequently in two passes. There are two possibilities – MAP after fMLLR and fMLLR after MAP. It can be anticipated

that the first approach should outperform the second one. The fMLLR method transforms all the mixtures from the same cluster at once, thus also mixtures with insufficient amount of adaptation data. MAP affects each of the mixtures separately, however only mixtures with sufficient amount of adaptation data are improved (shifted towards adaptation data – see equations (5)–(7)). Thus, the second pass (MAP adaptation) can be thought of as a refinement stage of mixtures with sufficient amount of data. In the case when fMLLR succeeds MAP, fMLLR (whole cluster is shifted at once) could disarrange mixtures that were correctly shifted by MAP (each mixture is shifted separately – more precise). This was also proved by the experiments shown in the second part of Table 1, therefore from now on we will focus on fMLLR followed by MAP adaptation. Note that in the case when huge amount of adaptation data would be available, the MAP adaptation would rearrange each of the mixtures, hence the fMLLR would take no effect. However, the first pass of fMLLR could cause more precise estimates of data statistics. On the other hand, having low amount of adaptation data MAP would be negligible and only fMLLR would take effect (see results of experiments depicted in Figure 1). This is a very natural behaviour of a combination of adaptation techniques.

The main disadvantage of such an approach is the need to compute the statistics defined in Section 2.1 twice. The procedure is as follows:

$$\text{SI} \rightarrow stats_1 \text{ for SI} \overset{\text{fMLLR}}{\Longrightarrow} \text{SD}_{\text{fMLLR}} \rightarrow stats_2 \text{ for SD}_{\text{fMLLR}} \overset{\text{MAP}}{\Longrightarrow} \text{SD}_{\text{fMLLR+MAP}} . \quad (18)$$

This approach brings an additional improvement of results (see Table 1), but it is not efficient with regard to the time consumption of computing the second statistics, e.g. in on-line adaptation. Hence, it would be suitable to adjust the already accumulated statistics without the need to see the feature vectors once again. Because fMLLR is in use, the adaptation stands for the transformation of feature vectors. Thus, instead of accumulating new statistics of transformed features, it is possible to transform the original statistics in the following way (consider expression (4) and the transformation $\bar{\boldsymbol{o}}_t = \boldsymbol{A}_{(n)}\boldsymbol{o}_t + \boldsymbol{b}_{(n)}$):

$$\bar{\varepsilon}_{jm}(\boldsymbol{o}) = \frac{\sum_{t=1}^{T} \gamma_{jm}(t)(\boldsymbol{A}_{(n)}\boldsymbol{o}(t) + \boldsymbol{b}_{(n)})}{\sum_{t=1}^{T} \gamma_{jm}(t)} = \boldsymbol{A}_{(n)}\varepsilon_{jm} + \boldsymbol{b}_{(n)} , \quad (19)$$

$$\begin{aligned} \bar{\varepsilon}_{jm}(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) &= \frac{\sum_{t=1}^{T} \gamma_{jm}(t)(\boldsymbol{A}_{(n)}\boldsymbol{o}(t) + \boldsymbol{b}_{(n)})(\boldsymbol{A}_{(n)}\boldsymbol{o}(t) + \boldsymbol{b}_{(n)})^{\mathrm{T}}}{\sum_{t=1}^{T} \gamma_{jm}(t)} = \\ &= \boldsymbol{A}_{(n)}\varepsilon_{jm}(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}})\boldsymbol{A}_{(n)}^{\mathrm{T}} + 2\boldsymbol{A}_{(n)}\varepsilon_{jm}(\boldsymbol{o})\boldsymbol{b}_{(n)}^{\mathrm{T}} + \boldsymbol{b}_{(n)}\boldsymbol{b}_{(n)}^{\mathrm{T}} , \end{aligned} \quad (20)$$

As can be seen from (19) and (20), the only approximation consists in the use of SI mixtures' posterior defined in (2), which remained unchanged (untransformed). The procedure of the proposed method is as follows:

$$\text{SI} \rightarrow stats_1 \text{ for SI} \overset{\text{fMLLR}}{\Longrightarrow} \text{SD}_{\text{fMLLR}} \rightarrow \text{transform } stats_1 \overset{\text{MAP}}{\Longrightarrow} \text{SD}_{\text{fMLLR+MAP}} . \quad (21)$$

Thus, data statistics $stats_1$ are computed according to the SI model. The fMLLR transformation matrices $\boldsymbol{A}_{(n)}$, $\boldsymbol{b}_{(n)}$ defined in (10) are estimated and utilized to transform the statistics $stats_1$ using equations (19), (20). At the end, the transformed statistics are used to refine the model mixtures via MAP adaptation.

3 Experiments

3.1 Test Data

All of the experiments were performed using telephone speech data set. The telephone-based corpus consists of Czech read speech transmitted over a telephone channel. The digitization of an input analog telephone signal was provided by a telephone interface board DIALOGIC D/21D at 8 kHz sample rate and converted to the mu-law 8 bit resolution. The corpus was divided into two parts, the training set and the testing set. The training set consisted of 100 speakers, where each of them read 40 different sentences (length of each sentence was cca 5 sec.). The testing set consisted of 100 speakers not included in the training set, where each of them read the same 20 sentences as the other, further divided into two groups. The first one contained 15 sentences used as adaptation data and the second one contained 5 remaining sentences used for testing of adapted models. The vocabulary in all our test tasks contained 475 different words. Since several words had multiple different phonetic transcriptions the final vocabulary consisted of 528 items. There were no OOV (Out Of Vocabulary) words. The basic speech unit of our system is a triphone. Each individual triphone is represented by a three states HMM; each state is provided by 8 mixtures of multivariate Gaussians. We are considering just diagonal covariance matrices. In all recognition experiments a language model based on zerograms was applied. It means that each word in the vocabulary is equally probable as a next word in the recognized utterance. For that reason the perplexity of the task was 528.

3.2 Results

The results of the experiment are shown in Table 1. The first part of the table contains the Correctness (Corr) and the Accuracy (Acc) of the baseline system (recognition done utilizing only the SI model) and the system adapted only by MAP or fMLLR method. Results of the adaptation process considering two-pass-combination (see (18)) of fMLLR and MAP can be found in the second part of the table. Last part of the table shows results of fMLLR+MAP combination proposed in this paper, where the data statistics were accumulated only once (for SI model) and then transformed using the estimated

Table 1. Correctness (Corr)[%] and Accuracy (Acc)[%] of transcribed words for each type of the adaptation and their combinations

	Corr[%]	Acc[%]
SI model	73.97	66.18
MAP	80.06	74.27
fMLLR	80.27	76.61
two-pass-combination		
MAP+fMLLR	81.91	77.68
fMLLR+MAP	83.05	79.51
one-pass-combination		
fMLLR+MAP	82.69	78.84

fMLLR transformation. At the end, the transformed statistics were used for MAP adaptation (see (21)).

As can be seen from Table 1, both two-pass-combinations MAP+fMLLR and fMLLR+MAP increse the Correctness (Corr) as well as the Accuracy (Acc) of recognized words, but the fMLLR+MAP approach is significantly better (as was supposed in Section 2.4). The result of the one-pass-combination of fMLLR+MAP is very close to the result of the two-pass-combination of fMLLR+MAP, which can be thought of as the upper boundary of the proposed method. We have performed also experiments based on increasing number of adaptation sentences, these can be found in Figure 1.

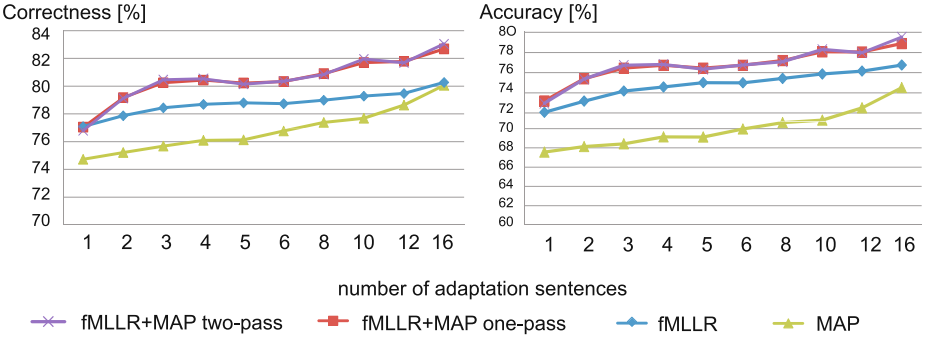

Fig. 1. Correctness (Corr)[%] and Accuracy (Acc)[%] of MAP and fMLLR adapted models and their combinations for increasing number of adaptation sentences. Note that the fMLLR+MAP combination yields the best performance for all different numbers of adaptation sentences and the results of one-pass-combination are very close (in some cases even identical) to the results of the two-pas-combination.

4 Conclusion

In this paper methods for adaptation of an acoustic model and their combinations were presented. More precisely, MAP and fMLLR adaptation methods were described. A simple combination of two different adaptation techniques proved to be convenient, but significantly more time consuming. Hence, we proposed a re-adjustment of primary computed data statistics without the need to see the adaptation data twice. We have demonstrated on experiments that such an one-pass-combination approach of fMLLR and MAP adaptation brings an additional improvement into the speech recognition. We have achieved a 4.57% and 2.23% increase absolutely in the systems' accuracy against the fMLLR and MAP based model, respectively. Let also state that the method proposed in this paper approaches the result of the two-pass-combination of fMLLR followed by MAP, which is however more time consuming.

Acknowledgements

This research was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/08/0707, by the Ministry of Education of the Czech Republic, project No.

MŠMT LC536 and by the Grant Agency of Academy of Sciences of the Czech Republic, project No. 1QS101470516.

References

1. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: Readings in speech recognition, pp. 267–296 (1990)
2. Psutka, J., Müller, L., Matoušek, J., Radová, V.: Mluvíme s počítačem česky, Academia, Praha (2007) ISBN:80-200-1309-1
3. Gauvain, L., Lee, C.H.: Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions SAP 2, 291–298 (1994)
4. Alexander, A.: Forensic Automatic Speaker Recognition using Bayesian Interpretation and Statistical Compensation for Mismatched Conditions. Ph.D. thesis in Computer Science and Engineering, pp. 27-29, Indian Institute of Technology, Madras (2005)
5. Leggeter, C.J., Woodland, P.C.: Maximum Likelihood Linear Regression for Speaker Adaption of Continuous Density Hidden Markov Models. Computer Speech and Language 9, 171–185 (1995)
6. Gales, M.J.F.: Maximum Likelihood Linear Transformation for HMM-based Speech Recognition. Tech. Report, CUED/FINFENG/TR291, Cambridge Univ. (1997)
7. Povey, D., Saon, G.: Feature and Model Space Speaker Adaptation with Full Covariance Gaussians. In: Interspeech, paper 2050-Tue2BuP.14 (2006)
8. Gales, M.J.F.: The Generation and use of Regression class Trees for MLLR Adaptation, Cambridge University Engineering Department (1996)
9. Machlica, L., Zajíc, Z., Pražák, A.: Methods of Unsupervised Adaptation in Online Speech Recognition. In: Specom, St. Petersburg (2009)