THE SPEAKER ADAPTATION OF AN ACOUSTIC MODEL

Lukáš MACHLICA [1], **Zbyněk ZAJÍC** [2]

Abstract: This paper deals with several adaptation techniques, which are of the importance in cases when the identity of a speaker is known and we want to recognize his speech. We are using three different methods, namely Maximum Apriori Probability adaptation, Maximum Likelihood Linear Regression and Constrained Maximum Likelihood Linear Regression. Each of the methods yields various benefits, therefore we examined their combination. This approach brought further error rate decreasing. All acoustic models are based on the Hiden Markov Model.

Keywords: adaptation, MAP, MLLR, CMLLR, combination.

1 INTRODUCTION

The Hidden Markov Model (HMM) with output probabilities described by Gaussian Mixture Models (GMM) has become an efficient tool for modeling of acoustic features in the speech recognition task in recent time [Rabiner (1990)]. To train the HMM, it is necessary to have large amount of data from many speakers. The final model, speaker independent (SI), is then able to recognize a speech from any speaker. When the speakers identity is known, we could acquire additional lowering of the error rate by using a model trained on the data from the particular speaker. Such a model is called the speaker dependent (SD) model. The main problem by the construction of the SD model is the need of large database of utterances from one speaker. This problem is often non-solvable in praxis. The solution is provided by adaptation techniques. In our case, it is the SI model transformation in terms of achieving the maximum probability for new data. The first part of the paper describes three types of adaptation, namely Maximum A Posteriori Probability (MAP), Maximum Likelihood Linear Regression (MLLR) and Constained MLLR (CMLLR). Because each of the methods works differently, the second part of the paper is devoted to their combination. The experiments and the data are described in the third part of the paper. The comparisons of error rates of the speech recognition with SI and SD model can also be found here. The results are promising, the adaptation techniques improve the recognition and lower the error rates by 8%.

2 ADAPTATION

The difference between the adaptation and ordinary training methods is the prior knowledge about the model parameters distribution, usually derived from the SI model [Psutka and col. (2007)]. The adaptation adjusts the model so that the probability of the adaptation data would be maximized. This is equivalent to

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

where $p(\lambda)$ stands for the prior information about the distribution of the vector λ containing model parameters, $O^i = \{o_1^i, o_2^i, \dots, o_T^i\}, i = 1, \dots, E$ is the sequence of feature vectors related to one speaker, λ^* is the best estimation of the SD model parameters.

The most relevant parameters, containing the information about the speaker, are means and covariance matrices of output probabilities of the HMM states represented by

[1] Ing. Lukáš Machlica, University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 333123456, e-mail: machlica@kky.zcu.cz

[2] Ing. Zbyněk Zajíc, University of West Bohemia in Pilsen, Faculty of Applied Sciences, Department of Cybernetics, Univerzitní 22, 306 14 Pilsen, tel.: +420 333123456, e-mail: zzajic@kky.zcu.cz

GMMs. These parameters are the sphere of our interest. Following equations are common for all of the adaptation techniques and we will refer to them in the consequent text. Let $\gamma_{jm}(t) = \frac{\omega_{jm} p(o(t)|jm)}{\sum_{m=1}^{M} \omega_{jm} p(o(t)|jm)}$ be the mixture posterior, ω_{jm}, μ_{jm} and σ^2_{jm} are the weight, the mean and the variance of the mixture m in the j-th state of the HMM, respectively. Let $c_{jm} = \sum_{t=1}^{Tj} \gamma_{jm}(t)$ be the soft count of mixture m and let the vector $\varepsilon_{jm}(o) = \frac{\sum_{t=1}^{Tj} \gamma_{jm}(t) o(t)}{\sum_{t=1}^{Tj} \gamma_{jm}(t)}$ be the average of features in frames which align to mixture m in the j-th state. Note: $\sigma^2_{jm} = diag(C_{jm})$ is the diagonal from the covariance matrix C_{jm}.

2.1 Maximum Aposteriori Probability (MAP) adaptation

MAP is based on the Bayes method for estimation of the acoustic model parameters, with the unit loss function [Gauvain, Lee (1994)]. 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

$$\bar{\omega}_{jm} = [\alpha_{jm} c_{jm}/T + (1 - \alpha_{jm})\omega_{jm}]\chi \;, \tag{2}$$

$$\bar{\mu}_{jm} = \alpha_{jm} \varepsilon_{jm}(o) + (1 - \alpha_{jm})\mu_{jm} \;, \tag{3}$$

$$\bar{C}_{jm} = \alpha_{jm} \varepsilon_{jm}(o \cdot o^T) + (1 - \alpha_{jm})(\sigma^2_{jm} + \mu_{jm}\mu^T_{jm}) - \bar{\mu}_{jm}\bar{\mu}^T_{jm} \;, \tag{4}$$

$$\alpha_{jm} = \frac{c_{jm}}{c_{jm} + \tau} \;, \tag{5}$$

where α_{jm} is the adaptation coefficient, which controls the balance between the old and new parameters using empirically determined parameter τ. χ is a normalization factor, which guarantees that all the new weights of the mixture for one state sum to unity. 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) [Alexander (2005)].

2.2 Maximul Likelihood Linear Regression (MLLR) adaptation

In contrast to MAP, where large amount of data is needed for each component, MLLR reduces the number of available model parameters via clustering (commonly used is regression tree) of similar components. The transformation matrix is the same for all of the parameters from the same cluster $K_n, n = 1, \dots, N$. Therefore MLLR needs less data and the adaptation is faster. The auxiliary function which has to be maximized takes the form [Gales (1997)]:

$$Q(\lambda, \bar{\lambda}) = const - \frac{1}{2} \sum_{jm} \sum_{t} \gamma_{jm}(t)(const_{jm} + \log|\bar{C}_{jm}| + (o(t) - \bar{\mu}_{jm})^T \bar{C}^{-1}_{jm}(o(t) - \bar{\mu}_{jm})) \,. \tag{6}$$

Mean is transformed according to the formula:

$$\bar{\mu}_{jm} = A_{(n)}\mu_{jm} + b_{(n)} = W_{(n)}\xi_{jm} \;, \tag{7}$$

where μ_{jm} is the original mean of the m-th mixture in the j-th state of the HMM, $\bar{\mu}_{jm}$ is the new adapted mean, $A_{(n)}$ is the regression matrix, $b_{(n)}$ is the additive vector, $\xi_{jm} = [\mu^T_{jm}, 1]^T$ is the original mean extended by 1 and $W_{(n)} = [A_{(n)}, b_{(n)}]$ is the transformation matrix for cluster K_n. Part of auxiliary function (6), which changes with the current transform $W_{(n)}$ is:

$$Q_{W_{(n)}} = const - \sum_{jm \in K_n} c_{jm} \sum_{i=1}^{I} \frac{(w^T_{(n)i}\xi_{jm})^2 - 2(w^T_{(n)i}\xi_{jm})\varepsilon(o)_{jm}(i)}{\sigma^2_{jm}(i)} \;, \tag{8}$$

where the column vector $w_{(n)i}$ is the transpose of the i-th row of $W_{(n)}$ and I is the dimension of feature vectors. Equation (8) can be further rearranged [Povey (2006)]:

$$Q_{W_{(n)}} = w^T_{(n)i}k_{(n)i} - 0.5w^T_{(n)i}G_{(n)i}w_{(n)i} \;, \tag{9}$$

where

$$k_{(n)i} = \sum_{jm\in K_n} \frac{c_{jm}\xi_{jm}\varepsilon(o)_{jm}(i)}{\sigma^2_{jm}(i)} \tag{10}$$

and

$$G_{(n)i} = \sum_{jm\in K_n} \frac{c_{jm}\xi_{jm}\xi^T_{jm}}{\sigma^2_{jm}(i)} \;. \tag{11}$$

And finally the maximum of equation (9):

$$\frac{\partial Q(\lambda,\bar{\lambda})}{\partial W_{(n)}} = 0 \Rightarrow w_{(n)i} = G^{-1}_{(n)i}k_{(n)i} \;. \tag{12}$$

2.3 fMLLR and CMLLR

In this case, compared to the MLLR, the transformation is applied on the feature space (feature MLLR). The auxiliary function changes to:

$$Q(\lambda,\bar{\lambda}) = const - \frac{1}{2}\sum_{jm}\sum_{t}\gamma_{jm}(t)(const_{jm} + \log|C_{jm}| - \log(|A_{(n)}|^2) + (\bar{o}(t)-\mu_{jm})^T C^{-1}_{jm}(\bar{o}(t)-\mu_{jm})). \tag{13}$$

The feature vectors are transformed instead of the model parameters, according to the formula [Ganitkevitch (2005)]:

$$\bar{o}(t) = A_{(n)}o(t) + b_{(n)} = A^{-1}_{(n)c}o(t) + A^{-1}_{(n)c}b_{(n)c} = W_{(n)}\xi(t) \;, \tag{14}$$

where $W_{(n)} = [A_{(n)}, b_{(n)}]$ is the transformation matrix, $\xi(t) = [o^T(t), 1]^T$ is extended feature vector and $A_{(n)c}, b_{(n)c}$ are matrices for equivalent transformation of parameters of the acoustic model:

$$\bar{\mu}_{jm} = A_{(n)c}\mu_{jm} - b_{(n)c} \;, \tag{15}$$

and

$$\bar{C}_{jm} = A_{(n)c}C_{jm}A^T_{(n)c} \;, \tag{16}$$

This method is called Constrained MLLR (CMLLR), because the same transformation matrix is used as for the means so for the covariances. In analogy with the previous section, it is possible to rearrange the auxiliary function (13) to the form [Povey (2006)]:

$$Q_{W_{(n)}}(\lambda,\bar{\lambda}) = \log(|A_{(n)}|) - \sum_{i=1}^{I} w^T_{(n)i}k_i - 0.5w^T_{(n)i}G_{(n)i}w_{(n)i} \;, \tag{17}$$

where

$$k_{(n)i} = \sum_{jm\in K_n} \frac{c_{jm}\mu_{jm}(i)\varepsilon(\xi)_{jm}}{\sigma^2_{jm}(i)} \;, \tag{18}$$

$$G_{(n)i} = \sum_{jm\in K_n} \frac{c_{jm}\varepsilon(\xi\xi^T)_{jm}}{\sigma^2_{jm}(i)} \;, \tag{19}$$

$$\varepsilon(\xi)_{jm} = [\varepsilon(x)_{jm}; 1] \;, \tag{20}$$

and

$$\varepsilon(\xi\xi^T)_{jm} = \left[\begin{array}{cc} \varepsilon(xx^T)_{jm} & \varepsilon(x)_{jm} \\ \varepsilon(x)_{jm}^T & 1 \end{array} \right] . \tag{21}$$

To find the solution of equation (17) we have to express $A_{(n)}$ in terms of $W_{(n)}$. It is possible to prove mathematically, that $\log(|A_{(n)}|) = \log(|w_{(n)i}^T c_{(n)i}|)$, where $c_{(n)i}$ is the cofactor of the matrix $A_{(n)}$ extended with a zero in the last dimension. After maximization of the auxiliary function (17) we receive:

$$w_{(n)i} = G_{(n)i}^{-1}(\frac{c_{(n)i}}{f} + k_{(n)i}) , \tag{22}$$

where

$$f_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} , \tag{23}$$

$$[a, b, c] = [\beta_{(n)}, -c_{(n)i}^T G_{(n)i}^{-1} k_{(n)i}, -c_{(n)i}^T G_{(n)i}^{-1} c_{(n)i}] , \tag{24}$$

$$\beta_{(n)} = \sum_{jm \in K_n} \sum_t \gamma_{(n)jm}(t) . \tag{25}$$

Because the equation (23) is a quadratic function, we obtain two different solutions $w_{(n)i}^{1,2}$. We choose the one, which maximalizes the auxiliary function (17). Subsequently we can compute the log likelihood for CMLLR as:

$$\log L(o(t)|\mu_{jm}, C_{jm}, A_{(n)c}, b_{(n)c}) = \log N(o(t); A_{(n)c}\mu_{jm} - b_{(n)c}, A_{(n)c} C_{jm} A_{(n)c}^T) , \tag{26}$$

or for fMLLR as:

$$\log L(o(t)|\mu_{jm}, C_{jm}, A_{(n)}, b_{(n)}) = \log N(A_{(n)} o(t) + b_{(n)}; \mu_{jm}, C_{jm}) + 0.5 \log(|A_{(n)}|^2) . \tag{27}$$

The estimation of $W_{(n)}$ is an iterative procedure, therefore matrices $A_{(n)}$ and $b_{(n)}$ have to be initialized first. The initialization for $A_{(n)}$ was chosen as a diagonal matrix with ones on the diagonal and initialization for $b_{(n)}$ as a zero vector. The estimation ends when the change in parameters of transformation matrices is small enough (about 20 iterations are sufficient).

2.4 Regression classes for MLLR

The benefit of the MLLR method is the possibility to cluster similar parameters (mixture components) of the model [Young and col. (2006)]. All parameters in a cluster are then transformed by the same transformation. The number of clusters depends on amount of adaption data. The parameters, which are close in the acoustic space, are clustered using Euclidean distance measure. Each of the final leaves of the tree contains just a single mixture component. The leaves are then merged together, until the root node (containing all the components from all mixtures) is obtained. The set of clusters in the regression tree is established during the adaptation process, according to the occupation count of each cluster with new data, where an empirical threshold has to be set. For example, lets have a look at the tree in the Figure 1 with four leaves $\{C_1, C_2, C_3, C_4\}$. The clusters C_3 and C_4 have a small occupation of adaptation data (lower then the threshold $T_{occup} = 850$). Therefore for all the components in clusters C_3 and C_4 will be used the transformation defined for the cluster C_5.

2.5 Combination of methods

Because each of the methods mentioned above works in different way, we have proposed the combination of the methods. This approach should increase the efficiency of adaptation. The first step is the SI model adaptation using MLLR. At the end we obtain the SD model$_1$. In the second step MAP adaptation is used for the SD model$_1$ to obtain the SD

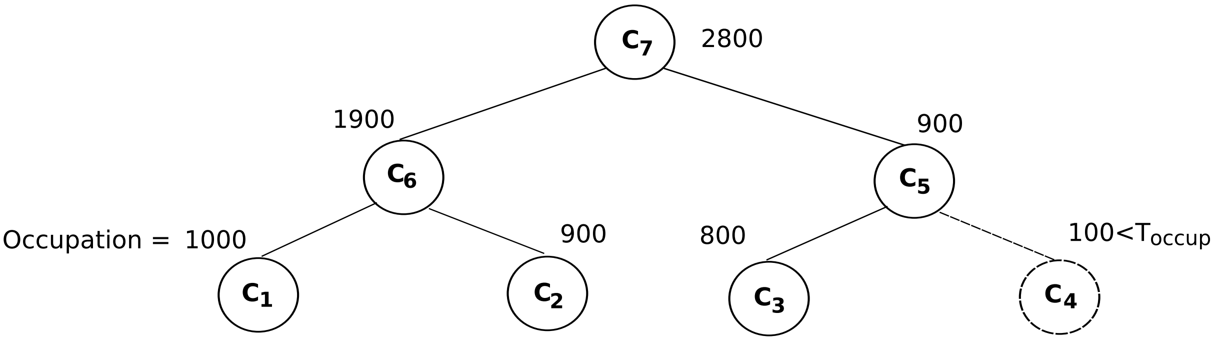

Fig. 1: Example of a binary regression tree. The numbers are the actual occupation counts of nodes (clusters). Nodes C_4 and C_3 have occupations lesser then the occupation threshold $T_{occup} = 850$ therefore for all the components of mixtures located in C_4 and C_3 will be used the transformation defined for node C_5.

model$_2$. In the case that there are not enough adaptation data, the MLLR method clusters similar mixture components and uses the same transformation for all the mixtures in the same group. In the second step, the MAP method refines the components of sufficiently occupied mixtures.

3 EXPERIMENTS

3.1 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. 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 monophone. Each individual monophone is represented by a three states HMM; each state is provided by 56 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

Table 1 shows results of the experiment. The baseline system (recognition done by the SI model) is in the first column. Another columns contain results obtained by MAP, MLLR and CMLLR, respectively. Last column shows combination of MAP and MLLR methods. We used the CMLLR transformation (not the fMLLR; the computation of transformation matrices is for both methods the same, the difference consists in evaluation of feature probabilities) and we considered just the diagonals of transformed covariance matrices. The MLLR results are superior to the CMLLR results. This can be explained by the fact, that constraints along with diagonal covariance matrices do not outperform the

SI model	MAP	MLLR	CMLLR	MLLR+MAP
21.27%	15.39%	15.98%	16.93%	13.67%

Tab. 1: WER[%] world eror rate .

unconstrained case of the mean adaptation by MLLR. Generally, fMLLR and CMLLR are equivalent transformations assuming that full covariance matrices are used. In comparison to the baseline system, all adaptation methods lower the World Error Rate (WER). The best performance is given by MAP adaptation due to sufficient amount of adaptation data. In the case, that the amount of adaptation data would be lesser, MLLR should outperform MAP. The results after combination prove that the mentioned methods have complementary information and WER falls further.

4 CONCLUSION

In this paper we have presented methods for adaptation of an acoustic model and their combination. We have described MAP, MLLR, CMLLR and combination of MLLR and MAP. We have demonstrated on experiments that the adaptation techniques bring a significant improvement. We have achieved WER reduction to 13.67% against the baseline system with WER 21.27%. In our future work we will aim at fMLLR instead of CMLLR and its extension to Speaker Adaptive Training (SAT) [Gales (1997)].

Acknowledgement: This research was supported by the Grant Agency of Academy of Sciences of the Czech Republic, project No.1ET101470416 and the Ministry of Education of the Czech Republic, project No. MŠMT LC536

REFERENCES

Alexander, A., 2005. Forensic automatic speaker recognition using bayesian interpretation and statistical compensation for mismatched conditions, *Ph.D. thesis in Computer Science and Engineering, Indian Institute of Technology, Madras, pp. 27-29.*

Gales, M.J.F., 1997. Maximum Likelihood Linear Transformation for HMM-based Speech Recognition, *Tech. Report, CUED/FINFENG/TR291, Cambridge Univ..*

Ganitkevitch, J., 2005. Speaker Adaptation using Maximum Likelihood Linear Regression. *Rheinish-Westflesche Technische Hochschule Aachen, the course of Automatic Speech Recognition, www-i6.informatik.rwth-aachen.de/web/Teaching/Seminars/SS05/ASR/Juri_Ganitkevitch_Ausarbeitung.pdf.*

Gauvain, L., Lee, C.H., 1994. Maximum a-posteriori estimation for multivariate Gaussian mixture observations of Markov chains, *IEEE Transactions SAP, 2:291–298.*

Povey, D., Saon, G., 2006. Feature and model space speaker adaptation with full covariance Gaussians, *Interspeech 2006, paper 2050-Tue2BuP.14.*

Psutka, J., Müller, L., Matoušek, J., Radová, V., 2007. Mluvíme s počítačem česky, *Academia, Praha 2007, ISBN:80-200-1309-1.*

Rabiner, L.R., 1990. A tutorial on hidden Markov models and selected applications in speech recognition, *Readings in speech recognition, pp. 267-296, ISBN:1-55860-124-4.*

Young, S., Evermann, G., Gales, M., and col., 2006. The HTK Book (for HTK 3.4). *User's manual, Cambridge University Engineering Department.*