Robust Statistic Estimates for Adaptation in the Task of Speech Recognition*

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@kky.zcu.cz, machlica@kky.zcu.cz, muller@kky.zcu.cz

Abstract. This paper deals with robust estimations of data statistics used for the adaptation. The statistics are accumulated before the adaptation process from available adaptation data. In general, only small amount of adaptation data is assumed. These data are often corrupted by noise, channel, they do not contain only clean speech. Also, when training Hidden Markov Models (HMM) several assumptions are made that could not have been fulfilled in the praxis, etc. Therefore, we described several techniques that aim to make the adaptation as robust as possible in order to increase the accuracy of the adapted system. One of the methods consists in initialization of the adaptation statistics in order to prevent ill-conditioned transformation matrices. Another problem arises when an acoustic feature is assigned to an improper HMM state even if the reference transcription is available. Such situations can occur because of the forced alignment process used to align frames to states. Thus, it is quite handy to accumulate data statistic utilizing only reliable frames (in the sense of data likelihood). We are focusing on Maximum Likelihood Linear Transformations and the experiments were performed utilizing the feature Maximum Likelihood Linear Regression (fMLLR). Experiments are aimed to describe the behavior of the system extended by proposed methods.

Keywords: fMLLR, adaptation, speech recognition, robustness.

1 Introduction

Nowadays, the adaptation of an acoustic model is used as a standard tool to improve the performance of Hidden Markov Model (HMM) in the task of speech recognition. In real conditions still several complications should be solved. The main problem arises in cases with low amount of adaptation data, even more if their quality is in question (noise, channel, etc.) [1]. Our effort is to decrease the vulnerability of the system in mentioned conditions.

In order to handle low amount of data the fMLLR adaptation is commonly utilized. Since fMLLR uses clustering of similar model parameters it suffices only with small

* This research was supported by the Ministry of Education of the Czech Republic, project No. MŠMT LC536, 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.

P. Sojka et al. (Eds.): TSD 2010, LNAI 6231, pp. 464–471, 2010.

amount of observation vectors. fMLLR approach is addressed in Section 2.2. However, in cases of on-line speech recognition [2], when data statistics are accumulated continuously, also fMLLR approach can be faced with difficulties in the sense of ill-conditioned transformation matrices.

The estimation of fMLLR matrices is an iterative procedure usually initialized with random matrices. Therefore, some procedures were developed to initialize fMLLR in order to ensure the stability of estimates of transformation matrices. One of the methods is described in Section 3.1. Another problem concerns inaccuracies in the phone alignment to HMM states caused by imperfect model training, and/or problematic data. To avoid the problem of incorrect model shifting via adaptation we are trying to discard non informative mixture components from the process of statistics accumulation (see Section 2.1). This approach, named Refinement of statistics, can be found in Section 3.2. Experiments were carried out utilizing two distinct corpora, one was used to adjust refinement values defined in Section 3.2, and the second was used to test the validity of these values, see Section 4. The article aims to describe the behavior of proposed methods, the discussion of results can be found in Section 4.4.

2 Adaptation Techniques

The task of adaptation is to shift the unadapted model in the direction of new (adaptation) data. The unadapted model is often denoted as Speaker Independent (SI) 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 mixture components, ω_{jm}, $\boldsymbol{\mu}_{jm}$ and $\boldsymbol{C}_{jm}$ are weight, mean and variance of the $m-th$ mixtures' component, respectively. Well know adaptation techniques are Maximum A-posteriori Probability (MAP) [3] and Linear Transformations based on the Maximum Likelihood [4].

2.1 Statistics of Adaptation Data

The adaptation techniques do not access the data directly, but only through some statistics defined as:

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

stands for the posterior of the $j-th$ state and the $m-th$ mixtures' component of the HMM. It should be noted that the pertinence of the feature vector $\boldsymbol{o}(t)$ to the $j-th$ state is given by the forced alignment process utilizing the reference transcription. Next,

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

is the soft count of mixture component m,

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

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

2.2 Feature Maximum Likelihood Linear Regression (fMLLR)

This technique belongs to the category of Linear Transformations (LTs), another LT based method is Maximum Likelihood Linear Regression (MLLR). These methods utilizes clustering of similar model components [6], thus clusters $K_n, n = 1, \ldots, N$ are formed. Hence, lower amount of adaptation data is needed to update the model. The fMLLR is used to transform directly 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 auxiliary function can be written in the form

$$Q_{\boldsymbol{W}_{(n)}}(\boldsymbol{\lambda}, \bar{\boldsymbol{\lambda}}) = \log|\boldsymbol{A}_{(n)}| - \sum_{i=1}^{I} \boldsymbol{w}_{(n)i}^{\mathrm{T}}\boldsymbol{k}_i - 0.5\boldsymbol{w}_{(n)i}^{\mathrm{T}}\boldsymbol{G}_{(n)i}\boldsymbol{w}_{(n)i} \ , \tag{6}$$

where

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

and

$$\boldsymbol{\varepsilon}_m(\boldsymbol{\xi}) = \left[\boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}), 1\right]^{\mathrm{T}} \ , \quad \boldsymbol{\varepsilon}_m(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}) = \begin{bmatrix} \boldsymbol{\varepsilon}_m(\boldsymbol{o}\boldsymbol{o}^{\mathrm{T}}) & \boldsymbol{\varepsilon}_m(\boldsymbol{o}) \\ \boldsymbol{\varepsilon}_m^{\mathrm{T}}(\boldsymbol{o}) & 1 \end{bmatrix} . \tag{8}$$

The solution of the minimization auxiliary function (6) can be found in [5]. Matrices $\boldsymbol{A}_{(n)}$ and $\boldsymbol{b}_{(n)}$ are estimated iteratively. Thus, they have to be initialized, e.g. as a diagonal matrix with ones on the diagonal and a zero vector, respectively.

3 Robustness

The task is to design the estimation of adaptation formulas as robust as possible in order to increase the systems efficiency in problematic situations - small amount of adaptation data, noise, model inaccuracies, etc. Such problems are handled in following subsections.

3.1 Inicialization of Adaptation Matrix

The auxiliary matrices (7) are dense and have a lot of parameters to be estimated. One of the problem arises in cases when low amount of adaptation data is available. Such situations can lead to ill-conditioned transformation matrices (5) and degradation of systems' performance. Therefore it is suitable to initialize matrices (7) with proper values in order to increase the robustness of the estimation process. The idea is to utilize data that fits the model to be adapted (when none new adaptation data are available, the estimated transformation matrix should equal the identity matrix). However, as already mentioned in Section 2.1 we do not need the data directly, we need only their statistics (mean and variance). Thus, we can use directly the unadapted model parameters as proposed in [8]. Now the initialization of (7) takes the form

$$\boldsymbol{k}_{(n)i} = \sum_{m \in K_n} p_m \frac{\mu_{mi}}{\sigma_{mi}^2} \begin{bmatrix} \boldsymbol{\mu}_m \\ 1 \end{bmatrix}, \; \boldsymbol{G}_{(n)i} = \sum_{m \in K_n} p_m \frac{1}{\sigma_{mi}^2} \begin{bmatrix} \boldsymbol{\mu}_m \boldsymbol{\mu}_m^T + \boldsymbol{C}_m & \boldsymbol{\mu}_m^T \\ \boldsymbol{\mu}_m & 1 \end{bmatrix}, \tag{9}$$

where $\boldsymbol{\mu}_m$,$\boldsymbol{C}_m$ are parameters of the m-th mixture component of the SI model belonging to the cluster $K_{(n)}$, p_m is a smoothing weight. Greater values of p_m indicate greater influence of the initialization, but the adaptation is less effective since the estimates are more restricted. In our case we set p_m equal to the weight of a mixture component, $p_m = \omega_m$. Other initialization approaches were studied in [9].

3.2 Refinement of Statistics

In order to accumulate the statistics (see Section 2.1) a proper phone alignment to HMM states is required. Even if a reference transcription is available the forced alignment can contain errors. These are caused among others by improper assumptions of the suitability of the HMM, e.g. the Maximum Likelihood (ML) training may not yield the most appropriate estimates [7].

We have investigated several approaches how to restrict the selection of mixture components used for estimation of transformation matrices. One of the possibilities is to discard a whole mixture component from a cluster (mixture component does not participate in (7)) based on its soft count (2). However, for some correctly aligned features the discarded component could be crucial (in the sense of its involvement in the estimation of G and k in (7)), but for some features such mixture component introduces only inaccuracies. Therefore it is more convenient to judge the components' suitability for each feature according to components' posterior defined in (1), and not reflect increments in statistic for inconvenient mixture components with low posteriors for a given feature (equivalent to setting the posterior $\gamma_{jm}(t)$ to zero in (7)).

Two possibilities were studied how to discard feature statistics according to the posterior $\gamma_{jm}(t)$. One could assign a threshold Th_γ and take into account only statistics related to mixture components with posteriors higher than the threshold Th_γ. Hence, $\gamma_{jm}(t) = 0$ if the inequality $\gamma_{jm}(t) \geq Th_\gamma$ is not met. Such an approach reflects an assumption that two hypothesis H_0, H_1 overlap, where H_1: *feature* o_t *was generated by mixture component* m, H_0: *feature* o_t *was NOT generated by mixture component* m, and we want to minimize the incorrect rejection of hypothesis H_0. A different approach accumulates only statistics acquired for N-best mixture components (with respect to their posterior) in one HMM state. These two techniques can be also combined - at first N-best mixture components are chosen and then the threshold Th_γ is applied to further exclude inconvenient components.

4 Experiments

We have utilized two data sets to test the systems performance. First set was used for estimation of refinement parameters and for some additional experiments related to amount of adaptation data and the second set was used to confirm the validity of estimated refinement values.

4.1 Czech Telephone (CzT) Corpus

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 1,260 different words. 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.

4.2 SpeechDat-East (SD-E) Corpus

SpeechDat-East [10] contains telephone speech in 5 languages Czech, Polish, Slovak, Hungarian, and Russian. For our experiments data were chosen in a similar fashion as CzT. We used only the Czech part of SD-E. The acoustic HMM was trained on 700 speakers with 50 sentences for each speaker (cca 5 sec. for sentence). For testing purposes 150 speakers were chosen with 50 sentences for each speaker. The vocabulary consisted of 7,000 words. No OOV words were present. Triphones were model using 3 state HMM with 8 gaussian mixtures (diagonal covariances) in each of the states. For the recognition a language model based on zerograms was considered.

4.3 Adaptation Setup

In our experiments we utilized fMLLR adaptation with clustering performed via the Regression Tree (RT). RT was constructed using the HTK toolkit [11]. The threshold for occupation of nodes in the regression tree was set to 1,000. Thus, approximately 10–15 matrices for one speaker were computed. Instead of the model directly the features were transformed. Only one iteration of fMLLR was carried out.

4.4 Results

First experiments involved the CzT set (see Section 4.1). Initially, we tried to find (empirically) the value of the threshold Th_γ for posterior $\gamma_{jm}(t)$ (see Section 3.2). Results of refined fMLLR (r-fMLLR) and combination of refined and initialized fMLLR (ri-fMLLR) can be found in Figure 1. As can be seen the best performance is achieved around the value $Th_\gamma = 0.5$ and a gain of 0.21% absolutely is acquired (see Table 1).

Instead of the threshold Th_γ one can choose N-best mixture components (according to their posterior $\gamma_{jm}(t)$ (1)) for statistics accumulation. Results of the N-best components approach together with results for $Th_\gamma = 0.5$ and basic fMLLR can be found in Table 1. These two refinement methods give very similar results. Basically, Th_γ controls the number of involved mixture components and so does the N-best approach. If Th_γ is set high enough only the best mixture component is chosen (though sometimes none) and the method works almost identically to the 1-best approach. The decrease of Th_γ is comparable to the increase of N. However, Th_γ balances the number of mixture components for each feature vector. Thus, in the rest of the paper we utilize only $Th_\gamma = 0.5$.

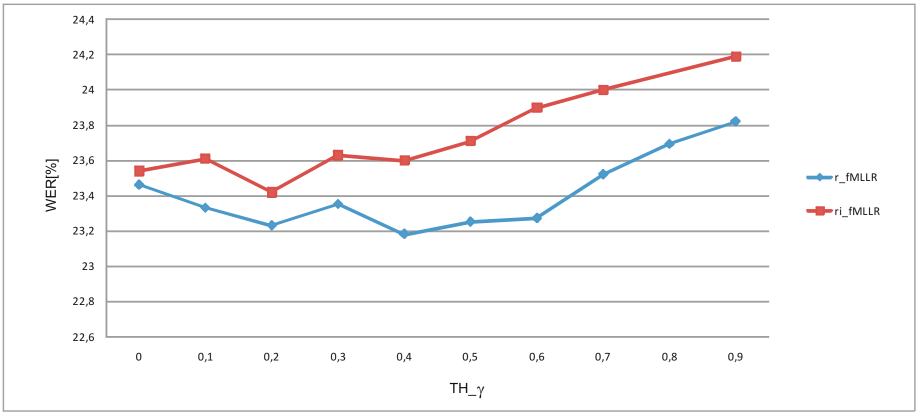

Fig. 1. Czech Telephone Corpus. Word Error Rate (WER)[%] of systems based on fMLLR adaptation with dependency on Th_γ, where r-fMLLR and ri-fMLLR denotes refined fMLLR and refined initialized fMLLR, respectively.

We have also investigated the dependency of the adaptation performance on the amount of adaptation data. We have compared the basic fMLLR approach with proposed improvements from Section 3. Results of basic fMLLR, i-fMLLR, r-fMLLR with $Th_\gamma = 0.5$, and their combination ri-fMLLR, with increasing number of adaptation sentences, are depicted in Figure 2. According to Figure 2, i-fMLLR outperforms the basic fMLLR, but better results are obtained for r-fMLLR. Combination ri-fMLLR performs worst. Since the threshold prunes the amount of data used for accumulation of statistics the initialization outweighs. Hence, the adapted model is more tightened to

Table 1. Czech Telephone Corpus. Word Error Rate (WER)[%] of unadapted system and refined fMLLR adapted system.

	WER[%]
unadapted system	33.82
basic fMLLR	23.46
r-fMLLR ($Th_{\gamma} = 0.5$)	23.25
$Nbest$-fMLLR ($N = 1$)	23.23
$Nbest$-fMLLR ($N = 2$)	23.33

the unadapted model. However, to preserve the robustness of the system (in the sense of well-conditioned fMLLR transformation matrices) it is useful to retain the initialization. Note that the robustly adapted system ri-fMLLR performs still better than basic fMLLR.

In order to prove the validity of refinement parameters we have carried out experiments also on the SD-E set described in Section 4.2 and results can be found in Table 2. Results proved the generality of refinement values estimated on the CzT set.

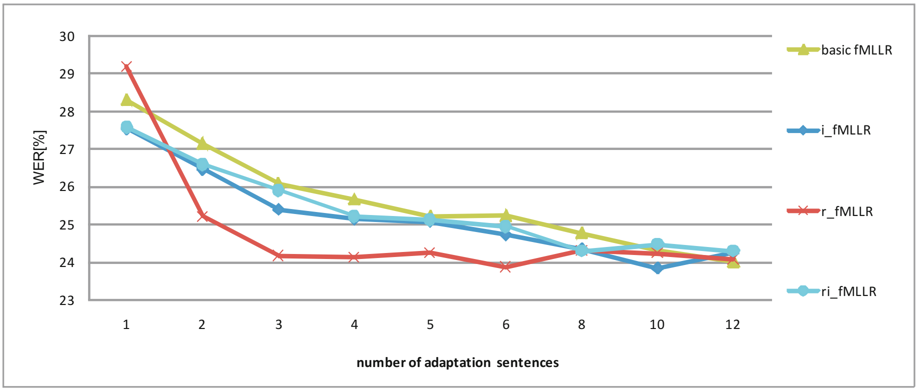

Fig. 2. Czech Telephone Corpus. Word Error Rate (WER)[%] of fMLLR adapted models for increasing number of adaptation sentences.

Table 2. SpeechDat-East Corpus. Word Error Rate (WER)[%] of unadapted system, refined fMLLR with $Th_{\gamma} = 0.5$ (r-fMLLR), initialized fMLLR (i-fMLLR) and combination of both (ri-fMLLR).

	WER[%]
unadapted system	55.18
basic fMLLR	46.12
r-fMLLR	44.50
i-fMLLR	44.66
ri-fMLLR	45.02

5 Conclusion

In this paper we have proposed procedures of robust estimation of adaptation parameters. We have investigated initialization of transformation matrices and refinement methods used to discard improper mixture components from the accumulation process of statistics. These techniques were applied to fMLLR approach, however they can be utilized also for any other estimations of linear transformations based on maximum likelihood. The main effort was dedicated to examination of the behavior of robustly adapted speech recognition system. Even if not all the proposed methods provided increase in systems performance, regarding the discussion in previous sections, the robustness of the system enhances. For example, such an approach is well suited for the task of adaptation in real-time recognition.

References

1. Psutka, J., Šmídl, L., Pražák, A.: Searching for a Robust MFCC-Based Parameterization for ASR Application. In: SIGMAP, Lisabon, pp. 196–199 (2007) ISBN: 978-989-8111-13-5
2. Pražák, A., Zajíc, Z., Machlica, L., Psutka, J.V.: Fast Speaker Adaptation in Automatic Online Subtitling. In: SIGMAP, Italy, pp. 126–130 (2009)
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. Gales, M.J.F.: Maximum Likelihood Linear Transformation for HMM-based Speech Recognition. Tech. Report, CUED/FINFENG/TR291, Cambridge Univ. (1997)
5. Povey, D., Saon, G.: Feature and Model Space Speaker Adaptation with Full Covariance Gaussians. In: Interspeech, paper 2050-Tue2BuP.14 (2006)
6. Gales, M. J. F.: The Generation and Use of Regression Class Trees for MLLR Adaptation. Cambridge University Engineering Department (1996)
7. Yu, K.: Adaptive Training for Large Vocabulary Continuous Speech Recognition. Ph.D. thesis, Hughes Hall College and Cambridge University Engineering Department (2006)
8. Li, Y., et al.: Incremental On-line Feature Space MLLR Adaptation for Telephony Speech Recognition. In: International Conference on Spoken Language Processing, Denver (2002)
9. Byrne, W., Gunawardana, A.: Discounted Likelihood Linear Regression for Rapid Adaptation. In: Eurospeech, Budapest, pp. 203–206 (1999)
10. Pollak, P., et al.: SpeechDat(E) – Eastern European Telephone Speech Databases, XLDB – Very Large Telephone Speech Databases. In: European Language Recources Association (ELRA), Paris (2000)
11. Young, S., et al.: The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department (2001-2006)