MAP Based Speaker Adaptation in Very Large Vocabulary Speech Recognition of Czech

Petr ČERVA, Jan NOUZA

Dept. of Electronics and Signal Processing, Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic

petr.cerva@vslib.cz, jan.nouza@vslib.cz

Abstract. *The paper deals with the problem of efficient adaptation of speech recognition systems to individual users. The goal is to achieve better performance in specific applications where one known speaker is expected. In our approach we adopt the MAP (Maximum A Posteriori) method for this purpose. The MAP based formulae for the adaptation of the HMM (Hidden Markov Model) parameters are described. Several alternative versions of this method have been implemented and experimentally verified in two areas, first in the isolated-word recognition (IWR) task and later also in the large vocabulary continuous speech recognition (LVCSR) system, both developed for the Czech language. The results show that the word error rate (WER) can be reduced by more than 20% for a speaker who provides tens of words (in case of IWR) or tens of sentences (in case of LVCSR) for the adaptation. Recently, we have used the described methods in the design of two practical applications: voice dictation to a PC and automatic transcription of radio and TV news.*

Keywords

Speech recognition, speaker adaptation, maximum a posteriori method, hidden Markov models.

1. Introduction

Modern systems for automatic speech recognition (ASR) are based on the technique that uses hidden Markov models (HMM), usually with continuous density function (CDHMM). Statistical parameters of these probabilistic models must be estimated in the phase of the system training by exploiting large databases of speech recordings.

In many practical applications, ASR systems are used by speakers whose speaking characteristics are different, depending on their gender, dialect, etc. To make the ASR robust against all these variations and to allow almost everybody to use the systems, these must operate as **speaker independent - SI**. For the training of a successful SI system, several tens of hours of annotated speech recorded by hundreds of different persons are necessary.

But there are applications, like e.g. voice control of a PC or voice dictation, where only one person is expected to use them. In such a case, the HMMs could be trained on the user's speech only and then the ASR would operate as **speaker dependent - SD**. Unfortunately, even in this specific case the user would be required to record at least several hours of his/her speech for training purposes. Generally this is hardly feasible for the user as well as for the provider of the system.

A more acceptable solution consists in the adaptation of the existing SI models for the given speaker and thus making the system **speaker adapted - SA**. The major advantage of this approach is that the speaker will be asked to record a significantly smaller amount of data (less than 1 hour). This recording and retraining is usually done on the user's own computer, which means that the ASR system also adapts its parameters to the characteristics of the given microphone and to specific noise of the environment in which the speaker talks. That is why adaptation methods are so important for practice and many researchers are still working on techniques that lead to further improvement in recognition while asking less adaptation data from the user.

In this paper we describe the method that utilizes the Maximum A Posteriori (MAP) strategy to the adaptation of phonemically based isolated-word recognition (IWR) and continuous-speech recognition (CSR) systems. Both have been developed for the Czech language and can operate with very large vocabularies (above 100,000 words).

This paper is structured as follows: In the next section we describe the theoretical background of the MAP based speaker adaptation technique. The results from quite extensive experimental evaluation are presented in section 3 (for the IWR task) and section 4 (for the CSR task).

2. MAP Based Speaker Adaptation

The adaptation (estimation) method MAP (Maximum A Posteriori) [1] is based on a different strategy than the Maximum Likelihood Estimation (MLE) [2] that is widely used during the standard HMM training.

While the MLE assumes the model parameters to be unknown but fixed, the MAP assumes these parameters to be random variables with known prior distributions. This information about prior distributions compensates the lack of adaptation data during the adaptation process.

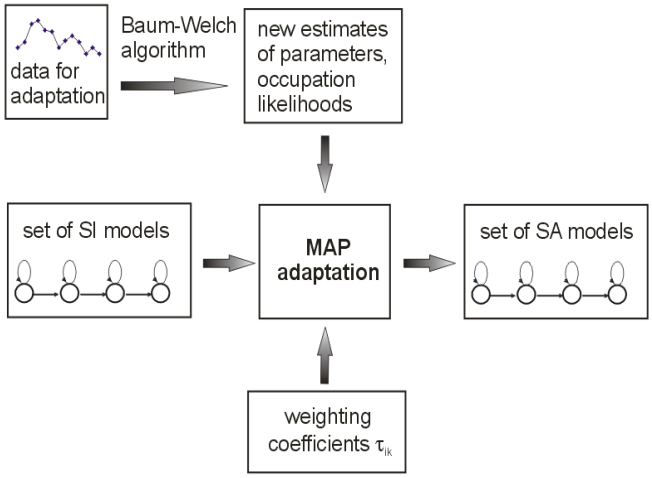

Fig. 1. Illustration of the MAP based adaptation process.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ be the sequence of parameterized vectors of adaptation data and $p(\Phi)$ the prior distribution of parameter Φ. The parameter Φ can represent means, variances or weighting coefficient of one Gaussian component (mixture) of the given HMM.

The optimal value of parameter Φaccording to MAP estimate can be expressed as

$$\Phi_{MAP} = \arg\max_{\Phi} p(\Phi|\mathbf{X}) \ , \tag{2.1}$$

where $p(\Phi|\mathbf{X})$ is the posterior probability density distribution of parameter Φ. According to Bayes' rule $p(\Phi|\mathbf{X})$ can be expressed as

$$p(\Phi|\mathbf{X}) = \frac{p(\mathbf{X}|\Phi)p(\Phi)}{p(\mathbf{X})} \ . \tag{2.2}$$

The maximization of $p(\Phi|\mathbf{X})$ is then reached by changing the value of Φ to maximize $p(\mathbf{X}|\Phi)\ p(\Phi)$.

The critical question in the MAP based speaker adaptation is: how to determine the prior parameters. In the standard MAP method these parameters are taken directly from the available SI models. In section 4.3 we will show that better results can be achieved if gender dependent SI models are used.

For CDHMMs with Gaussian state observation densities, where $\lambda_{ik} = (\mu_{ik}, \Sigma_{ik})$ is the k-th Gaussian component of state i with mixture weight c_k, the solution of the MAP estimation for the means is:

$$\boldsymbol{\mu}_{ik}^{SA} = \frac{\tau_{ik}}{\tau_{ik} + \sum_{t=1}^{T} \zeta_t(i,k)} \boldsymbol{\mu}_{ik}^{SI} + \frac{\sum_{t=1}^{T} \zeta_t(i,k)}{\tau_{ik} + \sum_{t=1}^{T} \zeta_t(i,k)} \hat{\boldsymbol{\mu}}_{ik} \ . \tag{2.3}$$

Here τ_{ik} is the weighting factor (a free parameter), $\boldsymbol{\mu}_{ik}^{SI}$ is the mean vector of k-th mixture component of state i of the given SI model, $\boldsymbol{\mu}_{ik}^{SA}$ is the adapted mean vector and $\hat{\boldsymbol{\mu}}_{ik}$ is the ML estimate of this vector computed by Baum-Welch algorithm [2] from all the adaptation data.

The term

$$\sum_{t=1}^{T} \zeta_t(i,k)$$

in (2.3) represents the occupation likelihood of k-th mixture component of state i and indicates the amount of the data used for the adaptation of this mixture.

Hence the new estimates of the SA parameters have form of weighted sum where the original SI parameters are weighted by factor

$$\tau_{ik} \Big/ \left(\tau_{ik} + \sum_{t=1}^{T} \zeta_t(i,k) \right)$$

and mixed with those estimated from the adaptation data by the MLE technique. This procedure is illustrated in Fig. 1.

When increasing the amount of the adaptation data

$$\lim_{t \to \infty} \sum_{t=1}^{T} \zeta_t(i,k) \to \infty$$

and the MAP estimate converges to the SD model which would result from the MLE method.

The formulae for the MAP estimation of the variances and mixture weights can be derived in a similar way. The expressions are rather complex and can be found in [3].

3. Experimental Evaluation in IWR Task

The evaluation was performed using the ASR system developed in our lab [5]. At the moment this is the most powerful IWR system designed for Czech. Its standard vocabulary contains 800,000 most frequent words. The system uses three-state CDHMMs of Czech phonemes [4] and noises. These are context independent, which is compensated by higher numbers of mixtures (32, 64 or even 100). The feature vector includes 39 MFCC parameters (13 static coefficients together with their first and second derivatives) calculated from the signal sampled at 8 kHz rate into 16 bits.

In the first series of experiments, we studied the influence of the amount of the adaptation data on the recognition accuracy of the adapted system (see Tab.1 and Fig.2). Here, only the Gaussian means were adapted using constant value τ_{ik} = 15 for all the mixtures. This value was found optimal in preliminary tests. The set of 431 Czech randomly chosen words was available for the adaptation, from which subsets with variable size were used in the adaptation procedure implemented in accord with the eq. (2.3). Another set of 862 words from the same speaker was used for the tests.

The tests were performed with two speakers and two system settings (32 or 64 mixtures) and the average results are given in Tab. 1.

number of adaptation words	0 (SI)	10	20	60	100	150	200	300	350	431
system with 32 mixtures										
recog. rate [%]	75	80	81	83	84	85	85	87	87	87
WERR [%]	–	21	24	30	35	40	40	46	45	47
system with 64 mixtures										
recog. rate [%]	74	81	82	83	85	85	85	87	87	88
WERR [%]	–	25	29	35	41	42	44	49	50	52

Tab. 1. Recognition accuracy as function of the number of adaptation words in IWR task

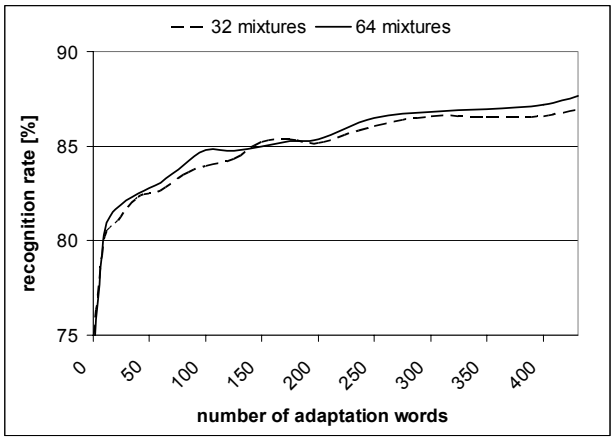

Fig. 2. Graph of the recognition accuracy as function of the number of adaptation words in IWR task.

We can notice that even if only 10 words are used for the adaptation, the relative word error rate reduction (WERR) is about 20 %. When 100 words are used the WERR reaches 35 %. With more adaptation data the improvement gets slower.

In another series of experiments, we studied the possibility to adapt also Gaussian variances and mixture weights. The results showed that the adaptation of these parameters had only a negligible effect on the recognition accuracy (see also section 4.2).

We also compared two types of adaptation data – isolated words vs. continuously spoken utterances and noticed that the former were more appropriate for the IWR task, most probably because of the voice stress which occurs at initial parts of isolated words.

4. Experimental Evaluation in CSR Task

In this case we used our own large vocabulary CSR system [7]. Its vocabulary was made of the 130,000 most frequent words. The acoustic part of the system was same as for the IWR task (64-mixture HMMs). The linguistic part was based on the bigram language model estimated on a corpus containing about 2 GB of Czech (mainly newspaper) text. Because the tests were time consuming, the initial experiments were performed only for one speaker. After finding the appropriate values of the free adaptation parameters we run another series of tests in which more speakers were involved.

4.1 Experiments Performed for One Speaker

Again, in the first series of experiments we studied the impact of the amount of the adaptation data – here measured by the number of adaptation sentences. The test set contained 462 sentences recorded by the same speaker. There were 7104 words in these test utterances, from which 225 were not present in the vocabulary, i.e. the Out-of-vocabulary (OOV) rate was 3.17 %. A varying number of other 0 to 600 sentences were used for the adaptation. The results are summarized in Tab. 2 and Fig. 3.

number of adaptation sentences	0 (SI)	25	50	75	100	200	400	955 (SD)
length [min]	–	3.0	6.4	9.2	12.4	25.3	51.0	122.1
recog. rate [%]	74.2	76.8	77.7	78.5	78.8	80.2	80.7	81.7
WERR [%]	–	10.2	13.6	16.9	18.0	23.3	25.4	29.3

Tab. 2. Recognition rate as function of the number of adaptation sentences for the LVCSR system.

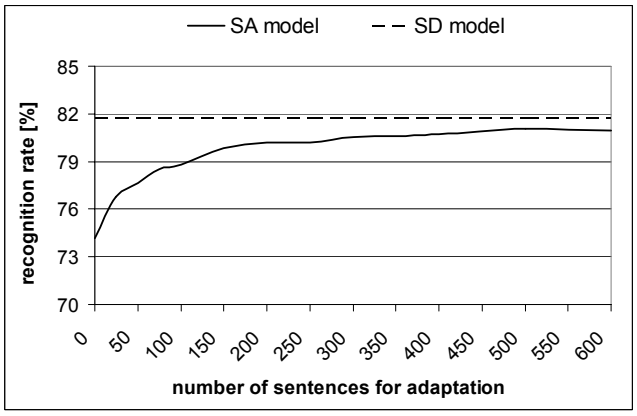

Fig. 3. Graph of the dependency of the recognition accuracy on the number of adaptation sentences for the LVCSR system with 64-mixture HMMs.

For comparison we created also SD models by using the total number of 955 sentences recorded by the same speaker. In Fig. 3 we can see the convergence of the SA system to the SD one as the amount of the adaptation data increases. The very important fact is that 100 sentences used in adaptation resulted in 18% WERR. In section 4.2 this significant improvement is verified for more speakers.

In the second series of experiments (see Tab. 3 and Fig. 4) we studied the impact of weighting coefficient τ_{ik}

on the recognition rate. Furthermore, also Gaussian variances and mixture weights were adapted in the same manner. Here, the adaptation set was fixed to 100 sentences.

weight τ_{ik}	1	2	3	4	5	8	10	15	20
adaptation of means									
recognition rate [%]	79.3	79.5	79.5	79.7	79.6	79.5	79.2	78.8	78.9
WERR [%]	19.7	20.4	20.5	21.1	21.1	20.5	19.3	17.8	18.1
adaptation of means and weights of mixtures									
WERR [%]	18.3	20.9	21.3	21.4	20.9	20.8	20.5	20.6	19.3
Adaptation of means, weights of mixtures and variances									
WERR [%]	20.0	19.9	20.5	21.5	21.7	21.3	20.5	21.0	20.0

Tab. 3. Results of MAP based adaptation for different values of adaptation weight and for different modes of adaptation.

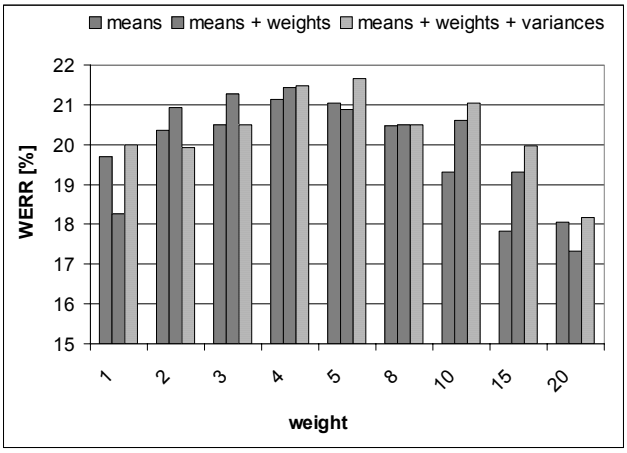

Fig. 4. Results of MAP based adaptation for different values of adaptation weight and for different modes of adaptation.

The results show that WERR values vary from 17.8 % to 21.1 % when the weighting coefficient τ_{ik} changes in range from 1 to 20 and when only means are adapted. This demonstrates the importance of the proper value of τ_{ik} on the results of the adaptation. The experiment also proves that the extension of MAP based speaker adaptation from the means only to the complete HMM parameters brings only a small absolute improvement in the WERR (about 0.6 % in average).

The optimal value of τ_{ik} depends on the amount of the adaptation data (see [3]) as well as on the type of adapted parameters. From Tab. 3 we can notice that for the given speaker the optimal value is close to 1 in case 100 adaptation sentences are used.

4.2 Experiments Performed for More Speakers

The recordings from seven other speakers were used in this evaluation. The people were either professional (broadcast) speakers or students. The former were recorded from radio news, the latter using standard microphone connected to a PC. The structure of this test database is shown in Tab. 4. In total there were 1566 sentences (more than 21,000 words) available for the testing and 100 sentences per speaker for the adaptation of HMM means. The weighting factor τ_{ik} was set to 1.

speaker	A	B	C	D	E	F	G
gender of the speaker	M	M	M	F	F	F	F
Microphone/ Radio	M	M	M	R	R	M	R
length of adapt. sentences [min]	12.4	15.5	12.8	7.5	7.5	7.5	7.5
SI system [%]	74.2	68.0	77.8	82.5	82.1	59.0	82.0
SA system [%]	79.3	73.0	81.9	86.2	85.7	68.0	84.4
WERR [%] (SA against SI)	19.8	31.1	18.5	21.1	20.1	22.0	13.3

Tab. 4. Results of MAP based adaptation of means for different speakers by using 100 adaptation sentences from each.

Tab. 4 proves that the adaptation was successful for all the speakers. The average reduction of word error rate (the value of WERR) was 20.8 %.

4.3 Application of Gender Dependent Models for Speaker Adaptation

It is known that male and female voices are different. This fact can be used also in speaker adaptation. Hence we performed another series of experiments with the same speakers as above and GD (gender dependent) models.

	speaker	A	B	C	D	E	F	G
SI models	recog. rate [%]	74.2	60.8	77.8	82.5	82.1	59.0	82.0
GD models	recog. rate [%]	76.6	61.2	79.3	84.4	83.8	60.0	84.5
	WERR [%]	9.3	1.0	6.8	10.9	9.5	2.4	13.9

Tab. 5. Speaker recognition rates achieved for 7 speakers and either general SI models or GD (gender dependent) ones.

The simplest adaptation technique only consisted in using gender dependent models trained either on male or female subsets of the general training database. As shown in Tab. 5 this simple approach led to small but consistent improvement (average WERR was 7.7 %).

In the second experiment the parameters of the GD models were used as priors for the MAP based adaptation of the means, i.e. the GD models rather than the SI ones were adapted. From Tab. 6 we can observe that using the prior GD models in the adaptation yielded slightly better results in most cases.

prior parameters	speaker	A	B	C	D	E	F	G
SI models	WERR [%]	19.8	31.1	18.5	21.1	20.1	22.0	13.3
GD models	WERR [%]	21.3	32.7	18.0	21.1	20.1	23.9	15.6
	Δ WERR [%]	1.5	1.6	-0.5	0.0	0.0	1.9	2.3

Tab. 6. Results of speaker adaptation performed by using GD and SI models as priors for the MAP based adaptation.

5. Conclusion

This paper deals with the problem of the efficient speaker adaptation for the very large vocabulary speech recognition of Czech. We focus on the MAP based methods whose theoretical background is briefly described. We applied this technique with several different variations, namely for Gaussian means, variances and mixture weights, using either gender independent or gender dependent models as prior information.

Our results from extensive experiments demonstrate that the word error rate (WER) in isolated-word recognition can be reduced by 20 % or even 35 % if the speaker provides 10 or 100 adaptation words, respectively. In the more complex task of continuous speech recognition, the WER will be reduced by about 20 % when 100 sentences are used for adaptation. Similar results have been reported also in [7], where they were achieved by commercial software (HTK). Our implementation has several advantages: It is more compact and flexible, because it allows further modification and optimization. One of them is the extension towards the adaptation of variances and mixture weights. In such case the WERR was further increased in average about 0.6 % in the task of LVCSR. Additional enhancement (about 1 %) was achieved by exploiting gender dependent models as priors within the MAP reestimation.

Recently, all the described methods are applied in the practical design of the systems for voice dictation into a PC. One version should serve as a Czech IWR dictation system working with 500,000-word vocabulary, the other as a CSR dictation machine with a 50,000 word lexicon. Thanks to the adaptation together with further improvement of lexical and language model, the recognition rate will get above the 90 % level. Another benefit from the research can be expected in the design of the system for the automatic transcription of broadcast news. The use of the special models adapted to the speech of key-speakers will improve the overall accuracy of the transcription.

Acknowledgments

This work has been partly supported by the Grant Agency of the Czech Republic (grant no. 102/02/0124) and through research goal project MSM 242200001.

References:

[1] GAUVAIN, J.L., LEE, C.H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. SAP*. 1994, vol. 2, p. 291 – 298.

[2] HUANG, X.D., ACERO, A., HON, H.W. *Spoken language processing*. Englewood Cliffs: Prentice Hall, 2001.

[3] ČERVA, P. Methods of speaker adaptation for speech recognition system. *Diploma thesis* (in Czech). TU of Liberec. 2004.

[4] NOUZA, J., PSUTKA, J., UHLÍŘ, J. Phonetic alphabet for speech recognition of Czech. *Radioengineering*. 1997, vol. 6, no. 4, p.16 to 20.

[5] NOUZA, J., NOUZA, T. A Voice dictation system for a million-word Czech vocabulary. *Proc. of Conference on Computing, Communication and Control Technologies*. Austin, 2004.

[6] NOUZA, J., NEJEDLOVA, D., ZDANSKY, J., KOLORENC, J. Very large vocabulary speech recognition system for automatic transcription of Czech broadcast programs. *Proc. of Int. Conference on Spoken Language Processing (ISCLP'04)*. Jeju, 2004.

[7] ŽELEZNÝ, M. Speaker adaptation in continuous speech recognition system of Czech. *PhD thesis* (in Czech). ZČU Plzeň 2001.

About Authors...

Jan NOUZA (1957) received master degree (1981) and doctor degree (1986) in telecommunications at the Czech Technical University (Faculty of Electrical Engineering) in Prague. Since 1986 he has been at the Technical University of Liberec (TUL). In 1999 he became professor at the Dept. of Electronics and Signal Processing. His major research field is speech interaction between human and computer with the special focus on speech recognition. He is the head of the Speech Processing Laboratory (SpeechLab) at the TUL founded by him in 1993. He is a IEEE member (Signal Processing Society) and a member of the International Speech Communication Association (ISCA).

Petr ČERVA was born in Liberec in 1980. In 2004 he received master degree at the Technical University of Liberec (TUL) and joined the SpeechLab team as a PhD student. His research work is focused on speaker adaptation and speech recognition of Czech.