

LONG TERM ON-LINE SPEAKER ADAPTATION FOR LARGE VOCABULARY DICTATION

Eric Thelen

Philips GmbH Forschungslaboratorien Aachen, Weißhausstr. 2, D-52066 Aachen, Germany
E-mail: thelen@pfa.research.philips.com

ABSTRACT

On-line speaker adaptation is desirable for speech recognition dictation applications, because it offers the possibility to improve the system with the speaker-specific data obtained from the user. Since the user will work with such a device over a long period, for a dictation system the long term adaptation performance is more important than the adaptation speed. In contrast to speaker-dependent re-training, the speaker-specific speech data does not need to be stored for on-line speaker adaptation and each adaptation step does not require a large computational effort.

In this paper we describe our way of performing on-line Bayesian speaker adaptation using partial traceback. We compare supervised with unsupervised adaptation and speaker adaptation with speaker-dependent training using the adaptation material.

Compared to the speaker-independent startup models, the error rate was divided by two after five hours of supervised adaptation in our experiments. In the long term experiments, supervised on-line adaptation performed similar to speaker-dependent training using the adaptation material.

1. INTRODUCTION

The importance of speaker adaptation methods for speech recognition dictation applications which need to achieve the speaker-dependent recognition performance without a separate user-specific training phase has led to an increasing research effort in this field. Two main approaches are the transformation of the incoming feature vectors (feature space techniques) and the adaptation of the parameters of the statistical acoustic models (model space techniques) [1]. The adaptation techniques depend on the structure of the acoustic models in the recognition system used. In our case, we use continuous mixture density hidden Markov models. For such models, Bayesian adaptation techniques are well established and have been successfully investigated [2]. Especially maximum a posteriori (MAP) adaptation, where the prior distributions can be derived from available models, has received much interest [3]. In this paper we focus on Bayesian speaker adaptation for the means of the probability density functions. We describe our way of performing on-line speaker

adaptation by making use of partial traceback and present results of long term experiments for large vocabulary dictation (≈ 20000 words).

We chose Bayesian adaptation for our work, because in a speech recognition dictation application [4], which will be used by a specific user for many hours, the long term adaptation performance is more important than the adaptation speed. Moreover, with a large amount of speaker-specific data available the performance of Bayesian adaptation should converge to the performance of speaker-dependent training [1, 2]. Other approaches, like maximum likelihood linear regression (MLLR [5]), primarily intend to achieve a faster adaptation process.

An unsupervised adaptation procedure is more convenient for a user-friendly dictation system, but the recognition errors may influence and mislead the adaptation process. In supervised adaptation, the adaptation process uses the text that has actually been spoken (recognition errors are corrected first). We compare our results for unsupervised and supervised speaker adaptation in this paper.

On-line adaptation means that the current model parameters are adapted whenever additional adaptation data is available and that the adaptation data does not need to be stored [6]. The computational effort for each incremental adaptation step is low compared to a speaker-dependent re-training.

Especially for long dictations we want to perform an on-line adaptation while the dictation still continues. In order to do this we make use of partial traceback which allows to obtain recognition results already during the recognition process. We describe the partial traceback procedure and the way we use it for on-line speaker adaptation. We compare the performance of speaker-adapted models with the performance of speaker-dependent ones trained on the adaptation material and comment on the speed of the adaptation process.

2. BAYESIAN SPEAKER ADAPTATION

In Bayesian speaker adaptation, background knowledge incorporated in (e.g. speaker-independent) reference models is combined with speaker-specific information collected during the adaptation process. According to [7], in Bayesian learn-

ing of the normally distributed mean ($p(\mu) \sim \mathcal{N}(\mu_o, \sigma_o^2)$) of a normal probability density function prior knowledge is exploited by

$$\mu_{new} = \frac{N\sigma_o^2}{N\sigma_o^2 + \sigma^2} m_{obs,ML} + \frac{\sigma^2}{N\sigma_o^2 + \sigma^2} \mu_o. \quad (1)$$

Interpreting equation (1) for the adaptation procedure means that N is the number of observations for the specific density in the adaptation material, $m_{obs,ML} = \frac{1}{N} \sum_{i=1}^N x_i$ is the mean of the adaptation observations (or the Maximum Likelihood (ML) estimate for the density mean [7]), σ^2 is the variance of the speaker-dependent observation generation process $p(x|\mu_{new}, \sigma^2)$ and μ_o and σ_o^2 are the parameters of the a priori distribution of the density mean. A possible estimate for the a priori mean μ_o is the mean that has been estimated in a (e.g. speaker-independent) previous training procedure. We set the (unknown) ratio of the variance σ^2 of the process $p(x|\mu_{new}, \sigma^2)$ to the a priori variance σ_o^2 to $\alpha = \sigma^2/\sigma_o^2$ and rewrite equation (1) as

$$\mu_{new} = \frac{N}{N + \alpha} m_{obs,ML} + \frac{\alpha}{N + \alpha} \mu_o. \quad (2)$$

This is the maximum a posteriori (MAP) estimate for the adapted mean of the probability density function. A large value of α decreases the influence of the adaptation observations leading to a slower but more stable adaptation process while a small value of α increases the influence of the adaptation observations with the risk of instabilities especially in the unsupervised adaptation procedure.

Another form of expressing equation (2) is the interpretation as a movement vector for the means of the densities, which leads to

$$\Delta\mu = \mu_{new} - \mu_o = \frac{N}{N + \alpha} (m_{obs,ML} - \mu_o). \quad (3)$$

If the number of adaptation observations increases ($N \rightarrow \infty$), we find $\mu_{new} \rightarrow m_{obs,ML}$, i.e. the MAP estimate converges to the ML estimate. If all observations of the adaptation material are assigned to the correct probability density function, the estimates of long term adaptation are identical to a speaker-dependent training using the adaptation material. It should be noted, however, that the density splitting procedure, which determines the distribution of mixture components in the acoustic space [8], is carried out only during the training process with the speaker independent data, so that there is still a difference between MAP adaptation and a complete speaker-dependent training. A possible effect is that not all mixture components are observed during the adaptation and thus the actual acoustic resolution is smaller than after a full speaker-dependent training.

Another difference to speaker-dependent training is that we only modify the means and not the other mixture parameters like mixture weights and variances.

3. ON-LINE ADAPTATION WITH PARTIAL TRACEBACK

As unsupervised adaptation methods suffer from the inaccuracy of the recognition results on the adaptation material,

it is important to use speaker-specific information for adaptation as soon as possible. For large vocabulary continuous speech dictations, we even want to perform adaptation steps while one dictation has not been completely recognized. In order to do this, we implemented a partial traceback procedure, which already provides first recognition results while the recognizer still processes the rest of the dictation.

3.1. Partial Traceback

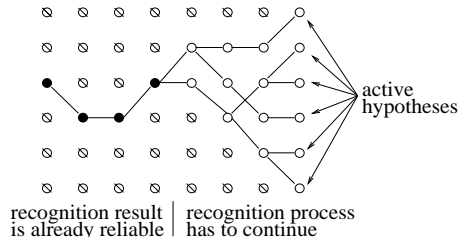


Figure 1: Partial Traceback

The partial traceback procedure is called after each N_{PT} frames. All the active hypotheses at the current point of the recognition process are considered. The procedure finds the latest point up to which all traceback paths of the active hypotheses are identical (Figure 1). All the words in the traceback history before this point have already been recognized and will not change during the further recognition process.

The adaptation procedure is supplied with these words and the affiliated frames. Note that in the unsupervised adaptation procedure no correction is carried out for the words or the word boundaries after the recognition result has been detected by the partial traceback. The supervised adaptation procedure replaces recognition errors with the words that have actually been spoken before each adaptation step.

3.2. On-line Adaptation

For our on-line adaptation procedure we use the information provided by the partial traceback as follows:

1. A Viterbi time alignment is performed for each of the recognized words (or each sequence of spoken words in the supervised adaptation procedure) using the affiliated frames resulting in the optimal state sequence for the respective observations. More specifically, each observation vector x is now assigned to a state s and thus to a mixture density m .
2. For each observation, we determine the density d of the mixture m with the largest influence in the likelihood computation for the observation vector x according to the maximum approximation [9].
3. In the adaptation step, we update the mean of each observed density d as follows:

$$\mu_{d,N+1} = \frac{x_{N+1} + (N + \alpha) * \mu_{d,N}}{N + 1 + \alpha} \quad (4)$$

Equation (4) is the recursive form of equation (2) with α and N having the same meaning. The recursion is initialized with

$\mu_{d,0} = \mu_o$. In order to ensure a minimum influence of each new observation, we limit N to 50 in the computation, i.e. after 50 observations for the density we do not increment N anymore. This can be seen as some kind of forgetting mechanism as suggested in [6], because all old observations and the startup models are downweighted with this limitation of N whenever more than 50 observations are available for a specific density.

After the described procedure, the adapted references are used while the combined recognition and adaptation process continues.

4. EXPERIMENTS

4.1. Experimental Setup

We evaluated the adaptation performance with our speech recognizer based on continuous Laplacian mixture density hidden Markov models. We use a single pooled deviation vector that is assumed to be independent of the densities and the states. The left-to-right hidden Markov models with three segments consist of two states with identical emission probability density functions [4]. The transition probabilities are not estimated (and not adapted) and given fixed values for loop, skip and forward transitions.

The long term adaptation experiments were performed with real-life speech data spoken by two male German speakers (Gm1, Gm2). The data consists of radiology dictations with an average duration of about 3 minutes. The speaker independent monophone references, which consist of 4054 densities, were trained with about one hour of speech data from each of eight different male German speakers.

After preliminary experiments on smaller test sets we decided to set the number of frames after which each partial traceback step is carried out to $N_{PT} = 3000$ (equivalent to 30 seconds of speech) and the adaptation parameter to $\alpha = 1/3$. The speech data for each of the two speakers was divided into adaptation and test material. During the recognition of the adaptation material the startup references are adapted on-line. Afterwards, the adapted references are evaluated with the independent test set.

The vocabulary size is 23271 words for Gm1 and 17547 words for Gm2. The test set perplexities (bigram language models) are 179 for Gm1 and 220 for Gm2.

4.2. Experimental Results

Tables 1 and 2 show the results of our experiments for Gm1 and Gm2, which are visualized in figures 2 and 3. The word error rates were measured on test sets of 29 minutes (Gm1) and 48 minutes (Gm2) which are independent of the adaptation material.

The word error rates obtained with the speaker-independent startup models are 21% and 31% for Gm1 and Gm2 respectively. The accumulated word error rates for the recognition of the adaptation material during the on-line adaptation are given in parentheses.

Table 1: Word error rates [%] for Gm1

adaptation data [min.]	unsupervised adaptation	supervised adaptation	speaker-dep. training
0	20.9 (SI)	20.9 (SI)	(100)
4	17.2 (31.3)	15.4 (29.9)	20.1
9	15.9 (32.4)	13.5 (29.9)	19.0
24	14.4 (29.6)	13.5 (26.0)	14.2
44	14.2 (27.4)	12.4 (24.6)	13.6
132	14.7 (26.8)	12.8 (24.2)	11.6
228	11.5 (24.6)	10.5 (22.2)	11.1

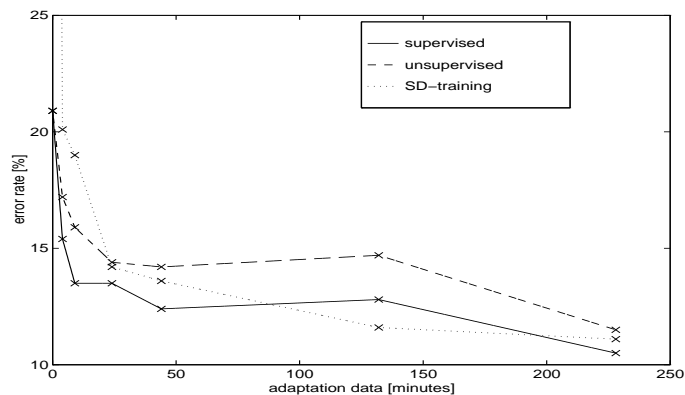


Figure 2: Long term adaptation for Gm1

For Gm1, 95% of all densities were adapted after 228 minutes and 98% of all densities were adapted after 497 minutes for Gm2.

Long Term Adaptation Performance (Adaptation vs. Speaker-Dependent Training)

For Gm1, the relative error rate reduction is 42% after 3.8 hours of unsupervised adaptation and 50% using supervised adaptation starting with the speaker-independent models. In the adaptation tests for Gm2, we obtained relative error rate reductions of 47% (unsupervised) and 54% (supervised) after 5 hours of adaptation.

The long term experiments show that supervised on-line adaptation performs similar to speaker-dependent training using the adaptation material.

We observed decreasing error rates with increasing amounts of speaker-specific data for the experiments with the independent test data except for a passage in the tests for Gm1 (between minute 44 and minute 132, see table 1). We assume that this part of the adaptation data is not representative for the test data and the chosen adaptation parameter $\alpha = 1/3$ is not optimal. However, the decreasing error rates for the speaker-dependent training and for the recognition of the adaptation data in the same passage indicate that the additional data can be used to compute more reliable estimates for the models.

After several hours of adaptation, there still seems to be a potential for further improvement with more speaker-specific

Table 2: Word error rates [%] for Gm2

adaptation data [min.]	unsupervised adaptation	supervised adaptation	speaker-dep. training
0	30.9 (SI)	30.9 (SI)	(100)
2	28.8 (34.1)	27.3 (26.7)	58.5
4	26.9 (30.5)	24.1 (25.4)	28.7
15	22.9 (28.9)	21.1 (20.4)	23.6
43	21.4 (22.5)	18.7 (17.4)	18.8
128	18.5 (17.4)	16.0 (15.0)	17.1
238	16.7 (15.6)	14.8 (13.8)	14.3
306	16.1 (16.2)	14.0 (14.2)	14.0
497	15.7 (15.7)	13.4 (13.8)	13.4

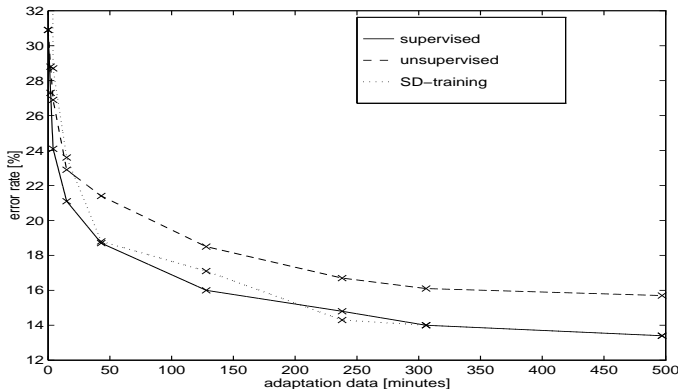


Figure 3: Long term adaptation for Gm2

data.

When only a small amount of adaptation material (less than 15 minutes) is available, the speaker-adapted models in the supervised and in the unsupervised experiments perform clearly better than both the speaker-independent ones and the speaker-dependent models trained on the adaptation material. With only little data, the speaker-dependent training cannot estimate reliable models. After about 20 minutes, the (supervised) speaker-dependent training provides generally better results than the unsupervised adaptation procedure while the supervised adaptation procedure still performs comparable to the speaker-dependent training.

Adaptation Speed

The relative improvement obtained with speaker adaptation is larger during the first minutes of the adaptation process. The adaptation speed slows down during the adaptation process.

Supervised vs. Unsupervised Adaptation

In all experiments, supervised adaptation clearly outperformed unsupervised adaptation. Regarding the absolute error rate, the distance between the performance of supervised and unsupervised adaptation already develops during the first minutes of the adaptation process and remains roughly constant during the long term adaptation.

5. CONCLUSIONS AND FUTURE WORK

Long term on-line speaker adaptation is useful for large vocabulary dictation applications, because compared to the straightforward Bayesian supervised adaptation algorithm that we used, speaker-dependent training does not achieve a better performance even when several hours of speaker-specific material are available. Supervised adaptation clearly outperforms unsupervised adaptation and should be used if the actually spoken text is accessible.

In our future work, we will investigate methods of adapting models that have not been observed in the adaptation material in order to accelerate the adaptation process.

6. ACKNOWLEDGEMENTS

The author thanks Ute Essen-Willemsen and Christian Dugast for their previous work in the field of speaker adaptation. Furthermore, he thanks Peter Beyerlein for providing the partial traceback procedure and Xavier Aubert, Hans-Guenter Meier and Bach-Hiep Tran for helpful discussions.

7. REFERENCES

1. L. Neumeyer, A. Sankar, V. Digalakis: "A Comparative Study of Speaker Adaptation Techniques", Proceedings EUROSPEECH, Madrid, Spain, September 1995, pp. 1127-1130
2. C.-H. Lee, C.-H. Lin, B.-H. Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", IEEE Transactions on Signal Processing, Vol. 39, No. 4, April 1991, pp. 806-814
3. C.-H. Lee, J.-L. Gauvain: "Speaker Adaptation Based on MAP Estimation of HMM Parameters", Proceedings ICASSP, Minneapolis, U.S.A., April 1993, pp. II558-II561
4. V. Steinbiss et al.: "Continuous Speech Dictation - From Theory to Practice", Speech Communication 17 (1995), pp.19-38
5. C.J. Leggetter, P.C. Woodland: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language (1995) 9, pp. 171-185
6. Q. Huo, C. Chan: "On-Line Bayes Adaptation of SCHMM Parameters for Speech Recognition", Proceedings ICASSP, Detroit, U.S.A., May 1995, pp. 708-711
7. R.O. Duda, P.E. Hart: "Pattern Classification and Scene Analysis", Wiley Interscience, New York, 1973
8. H. Ney: "Acoustic Modelling of Phoneme Units for Continuous Speech Recognition", Proceedings Fifth European Signal Processing Conference, Barcelona, Spain, September 1990, pp. 65-72
9. H. Ney: "Modeling and Search in Continuous Speech Recognition", Proceedings EUROSPEECH, Berlin, Germany, September 1993, pp. 491-498