A COMPARATIVE STUDY OF MODEL-BASED ADAPTATION TECHNIQUES FOR A COMPACT SPEECH RECOGNIZER

Frank Thiele, Rolf Bippus

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
Frank.O.Thiele, Rolf.Bippus@philips.com

ABSTRACT

Many techniques for speaker adaptation have been successfully applied to automatic speech recognition. This paper compares the performance of several adaptation methods with respect to their memory need and processing demand. For adaptation of a compact acoustic model with 4k densities, Eigenvoices and structural MAP (SMAP) are investigated next to the well-known techniques of MAP and MLLR adaptation. Experimental results are reported for unsupervised on-line adaptation on different amounts of adaptation data ranging from 4 to 500 words per speaker.

The results show that for small amounts of adaptation data it might be more efficient to employ a larger baseline acoustic model without adaptation. Eigenvoices achieve the lowest word error rates of all adaptation techniques but SMAP presents a good compromise between memory requirement and accuracy.

1. INTRODUCTION

Using state of the art speech recognition technology, speaker dependent speech recognition by far outperforms speaker independent recognition and whenever possible speaker adaptation techniques are used to close the gap.

However, speaker adaptation imposes additional effort on an application with respect to both memory and computational requirements. This is a severe problem especially for small command and control recognizers for which both resources are limited. A second consideration has to be the way speaker adaptation may be applied in a command and control recognizer. In contrast to e.g. a dictation application, speaker dependent sessions with a long prior enrollment phase are often not feasible. A very short enrollment and/or fast on-line adaptation are necessary, both requiring adaptation techniques that work on very little adaptation data.

Looking into suitable adaptation techniques, we may in general distinguish feature-based adaptation techniques that aim at modification or normalization of the recognizer input from model-based techniques that modify parameters of the acoustic model. Vocal tract length normalization [1] and the use of all-pass transforms [2] should be mentioned as prominent methods for the first category.

Focusing on model-based techniques in this paper, we further distinguish methods that are based on model selection and those that in general use a transformation to modify acoustic parameters.

Selection-based methods use a finite set of acoustic models for different speaker clusters and adaptation itself is performed by either selecting the best fitting set of parameters or by interpolating multiple best sets [3]. Gender dependent modeling [4] is the most simple but widely used example of such a method. Using a selection-based method in a speaker independent on-line system, speaker consistent parsing [5] may be used. Especially for small systems, the resulting memory and computational effort is however equivalent to simply increasing the number of modeling parameters within the mixtures of the acoustic model, a method that outperformed model selection (gender dependent and speaker cluster dependent modeling) in experiments we previously performed on the data used in this paper.

Our focus therefore lies on transformation-based adaptation, starting with standard MAP [6] and MLLR [7] adaptation. The literature reports a number of techniques to meet the severe performance degradation these methods suffer for small amounts of adaptation data. As a first step tree-organized regression classes are used in MLLR [7], however it still suffers from the fact that the ML paradigm is used, not taking prior knowledge into account. Even when using regression classes, for many classes insufficient or no data will be present for a robust estimation of transformation parameters. The literature therefore suggests a number of methods to explore the dependency of modeling parameters using prior knowledge to allow for adaptation of unseen parameters by predicting their values from the observed ones [8, 9].

Here we explore Eigenvoices and Structural Maximum A Posteriori (SMAP) adaptation.

Eigenvoice adaptation [9] uses a principal component analysis in the parameter space to obtain principal directions of parameter changes to which parameter adaptation is restricted. An extension of the Eigenvoice approach to general MAP adaptation as we use it is given in [10]. Structural Maximum A Posteriori (SMAP) adaptation applies a tree structure on the densities to describe the dependencies between density parameters. SMAP was first proposed by Shinoda and Lee in 1997. A detailed description by the authors can be found in [11]. An extension of SMAP to SMAPLR (not used in this paper) using linear transformations (as in MLLR) is presented in [12].

The following section describes the adaptation methods as applied for the experiments presented in the sections thereafter.

2. ADAPTATION TECHNIQUES

This section gives a short overview of the investigated adaptation techniques with respect to their memory requirements.

In this paper we focus on adaptation of the means of CDHMMs with one globally tied diagonal covariance matrix using Viterbi alignment on the best recognized sentence hypothesis. Let

N be the total number of mean vectors $\boldsymbol{\mu}_d, d = 1, \ldots, N$, with dimension D, of the baseline acoustic model.

2.1. MLLR

In maximum likelihood linear regression (MLLR) [7], a density mean $\boldsymbol{\mu}_d$ is linearly transformed like

$$\widehat{\boldsymbol{\mu}}_d = A\boldsymbol{\mu}_d + \boldsymbol{b} =: T\tilde{\boldsymbol{\mu}}_d, \tag{1}$$

where $\tilde{\boldsymbol{\mu}}_d$ is the augmented mean with dimension D+1. Maximum likelihood estimation reduces to the least squares estimate: $T = M_{\mathrm{o}} M_{\mathrm{m}}^{-1}$ with

$$M_{\mathrm{o}} = \sum_t \boldsymbol{o}_t \tilde{\boldsymbol{\mu}}^T_{\mathrm{density}(t)}, \tag{2}$$

$$M_{\mathrm{m}} = \sum_t \tilde{\boldsymbol{\mu}}_{\mathrm{density}(t)} \tilde{\boldsymbol{\mu}}^T_{\mathrm{density}(t)}. \tag{3}$$

The three matrices $M_{\mathrm{o}}, M_{\mathrm{m}}$, and T require storage of $(D + 1)(2.5D+1)$ floats.

For MLLR with multiple regression classes, we need to store a matrix $M_{\mathrm{o},r}$ for each class r and a class index for each density. If indices of the observed densities are stored, only one matrix M_m is needed and can be calculated during the adaptation step. Allowing to collect 60 seconds of adaptation data at a frame rate of 16ms, 3750 density indices have to be kept for each adaptation step. Regression class tying requires storage of the class tree and 2 additional matrices for temporary storage of $M_{\mathrm{o}}, M_{\mathrm{m}}$.

The computational demand of MLLR adaptation is moderate. For each regression class, the symmetric $(D+1)\times(D+1)$ matrix M_{m} has to be inverted.

2.2. MAP

For the system considered here, maximum a posteriori (MAP) adaptation [6] boils down to a linear interpolation of observed and prior mean for each density:

$$\widehat{\boldsymbol{\mu}}_d = \boldsymbol{\mu}_d + \frac{N_d}{N_d + \tau}(\boldsymbol{\mu}_{\mathrm{obs},d} - \boldsymbol{\mu}_d), \tag{4}$$

where N_d is the the number of observations of density d, and τ a parameter describing the precision of the prior mean. For MAP adaptation, all observations are accumulated on density level. Therefore, a complete set of density means and observation counts has to be stored in addition to the current and the initial acoustic model.

The computational demand for the actual MAP adaptation is very small since the adapted density means are simply obtained as weighted sum.

2.3. Eigenvoices

Eigenvoice adaptation [9] is performed on the *ND* dimensional parameter vector $\underline{\boldsymbol{\mu}}$, the concatenation of all means $\boldsymbol{\mu}_d$. Basically, the parameter vector is moved along special directions $\underline{E}_i$ (Eigenvoices) which represent *a priori* knowledge about speaker variability:

$$\underline{\widehat{\boldsymbol{\mu}}} = \underline{\boldsymbol{\mu}} + \sum_{i=1}^{n} c_i \underline{E}_i. \tag{5}$$

Eigenvoice adaptation can be embedded in an anisotropic MAP framework [10] where adaptation into all directions is allowed but Eigenvoice directions are preferred.

For Eigenvoice adaptation only very few parameters need to be estimated, namely one coefficient c_i for each Eigenvoice. However, the estimation of these coefficients is rather expensive. In addition to the density accumulation as needed for MAP (sec. 2.2), the Eigenvoices have to be stored or loaded. Since each Eigenvoice is a complete set of density means, n Eigenvoices require nND float values.

To compute the Eigenvoice coefficients, a $n \times n$ matrix B has to be calculated and inverted. The calculation of B requires about n^2ND operations. If only the m most often observed densities are used to determine B, it is possible to reduce the amount of operations. This is exact as long as no more than m different densities were observed, i.e. for little adaptation data.

2.4. SMAP

Structural (hierarchical) MAP (SMAP) adaptation [11] exploits prior knowledge about the model parameters in the form of hierarchical organization of densities.

Cluster mismatch

In SMAP adaptation the mismatch between training and observation data is explicitly modeled. We assume that a cluster $\mathbb{M} = \{d_0, d_1, \ldots\}$ of densities is given. The maximum likelihood mismatch estimate $\boldsymbol{\nu}_{ML}$ for the cluster is given by the observations as

$$\boldsymbol{\nu}_{ML} = \frac{\sum_{d\in\mathbb{M}} N_d(\boldsymbol{\mu}_{\mathrm{obs},d} - \boldsymbol{\mu}_d)}{\sum_{d\in\mathbb{M}} N_d}. \tag{6}$$

The MAP mismatch estimate $\widehat{\boldsymbol{\nu}}$ for the cluster reduces to a linear interpolation of the observed mismatch $\boldsymbol{\nu}_{ML}$ and a prior mismatch $\boldsymbol{\nu}_0$:

$$\widehat{\boldsymbol{\nu}} = \frac{M}{M+\tau}\boldsymbol{\nu}_{ML} + \frac{\tau}{M+\tau}\boldsymbol{\nu}_0, \tag{7}$$

where $M = \sum_{d\in\mathbb{M}} N_d$ is the total number of observations of the cluster and τ is the precision of $\boldsymbol{\nu}_0$.

Each density mean in the cluster is then adapted to compensate for the mismatch:

$$\widehat{\boldsymbol{\mu}}_d = \boldsymbol{\mu}_d + \widehat{\boldsymbol{\nu}}, \quad \forall d \in \mathbb{M}. \tag{8}$$

Usually no prior information about the mismatch between the speaker independent model and the observed speaker (e.g. gender) is given and $\boldsymbol{\nu}_0 = 0$. For individual densities (no clustering), equations (7) and (8) then become the standard MAP adaptation (4).

Hierarchical priors

Now the densities are assumed to be embedded in a tree structure, where each node represents a density cluster. The basic idea of SMAP adaptation is to take the mismatch estimate at a node as the prior mismatch for all children nodes. For a node s of the tree, the mismatch estimate is then given as (cf. equation (7))

$$\widehat{\boldsymbol{\nu}}(s) = \frac{M(s)}{M(s)+\tau(s)}\boldsymbol{\nu}_{ML}(s) + \frac{\tau(s)}{M(s)+\tau(s)}\widehat{\boldsymbol{\nu}}(p(s)), \tag{9}$$

where $p(s)$ denotes the parent node of node s. For the root node, the prior mismatch is set to zero.

The mismatch estimate of a density is thus obtained as the weighted average of all ML estimates along the path from the root node to the respective leaf node. The weights are determined by the number of observations and by the $\tau(s)$.

Theoretically, $\tau(s)$ is the precision of the prior estimate $\widehat{\boldsymbol{\nu}}(p(s))$. As in MAP adaptation $\tau(s)$ is taken to be an empirical parameter allowing to control the weight of the ML estimates at different tree layers.

Memory need

For each tree node a mismatch vector and the observation count have to be stored. The tree structure itself requires pointers to the children and parent nodes. In addition to the density accumulation as needed for MAP, a tree with 3 children per node and $N+S$ nodes (N leaves) requires $S(D+1)$ floats and $N+4S$ pointers, which is less than 1 Eigenvoice.

Tree clustering

A suitable tree organization of the densities is essential for SMAP adaptation. The root node corresponds to the "cluster" of all densities and the children nodes then successively describe density clusters with less densities for each tree level. To obtain the full effect of MAP adaptation for large amounts of adaptation data, each leaf node corresponds to one density.

For density clustering we choose a bottom-up approach. A variance criterion based on Euclidean distances is used as distance measure: The two clusters with the smallest increase in heterogeneity are merged. After a specified number of successive merges of the closest clusters, k-means iterations reassign each density to the closest cluster.

For straightforward SMAP adaptation with depth-depending adaptation parameters τ a tree with fixed number of levels is constructed. Different trees can be constructed by varying the number of levels and the branching. Each node has a specified number b of branches (i.e. children), provided there are at least b densities at that node. At the lowest tree level above the leaves branching might be greater than b to obtain exactly one leaf for each density.

To organize density clusters in a tree with a given number of levels l and branching b, the tree structure is constructed top-down.

3. EXPERIMENTAL SETUP

For experiments we use an internal database with 162 mainly non-native English speakers recorded in an office environment at 8kHz. The number of males, females, boys and girls is roughly balanced. The database is split into 97 speakers for training and 65 for evaluation with about 500 words (235 utterances) per test speaker. The evaluation vocabulary of 152 words comprises numerical strings and command words for audio and TV control.

For feature extraction the samples are grouped into frames of 32 ms width at intervals of 16 ms. After extracting 12 mel-cepstral features, 8 delta features are added. We model 600 triphones with 2600 HMM states, 23 000 Laplacian densities are trained and clustered down to 4096 densities. One diagonal covariance matrix is shared by all densities. For comparison larger acoustic models with 8k and 16k densities (clustered from 35 000) are trained. Recognition is performed without network or language model, we simply apply a word penalty.

Unsupervised on-line adaptation is carried out on different amounts of adaptation data ranging from 2 to all 235 utterances per speaker. Results for n utterances are averaged over $235/n$ evaluations. Silence densities are not adapted.

The actual adaptation is always performed on the initial acoustic model using all accumulated observations. Switching on adaptation thus always means that two sets of means need to be stored: the initial and the current one which is used for decoding. From the point of view of memory consumption, an adaptation method should always be as accurate as the 8k baseline. Table 1 shows the memory need to store all means and density links of the acoustic models.

MLLR adaptation is performed with one class and 40 classes (regression class tying). Eigenvoices were obtained by PCA of all 97 training speakers. For SMAP a density tree with a branching of 3 and 8 layers (including root) was grown, the lowest 4 levels are used for adaptation.

ac. model	#densities	memory need	WER [%]
4k	4096	500 kB	12.2 ± 0.4
8k	8192	910 kB	10.7 ± 0.3
adaptation	4096	820 kB	–

Table 1: Memory demand of acoustic models and on-line adaptation.

4. RESULTS

Figure 1 shows the performance of the adaptation techniques. For 4 words of adaptation data Eigenvoices already perform significantly better than the 4k model (table 1) but cannot beat the 8k model. Even 1 Eigenvoice outperforms MAP, MLLR, and SMAP. Adapting on 10–40 words, MAP and MLLR still show no significant improvement while SMAP achieves a relative gain of up to 7% and Eigenvoices even of 17%. For adaptation on 200 words, all methods except for MAP are able to outperform the 8k model. Note that 500 words correspond to only 5 minutes of speech. The results of MAP, SMAP, and Eigenvoice adaptation should all converge asymptotically to the same (i.e. ML) estimate.

SMAP always gives better results than MAP and MLLR. It also achieves better performance than one Eigenvoice if enough adaptation data is available. Better SMAP results for small amounts of adaptation data have been reported in [11]. This might be due to a more suitable density tree or a stronger mismatch between training and evaluation.

Figures 2 and 3 show the performance of the adaptation techniques with respect to memory need for adaptation on 500 and 20 words respectively[1]. While for 500 words all adaptation methods easily outperform the unadapted acoustic models, for 20 words no method can beat the unadapted models at the same memory demand. It can be seen that Eigenvoices need the most memory by

[1]The memory requirements are estimated theoretically, assuming 4 bytes for a float value, 8 bytes for a pointer.

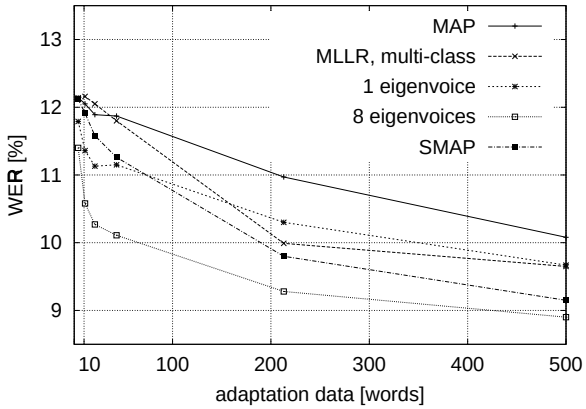

Figure 1: Word error rates for different amounts of adaptation data.

far to achieve their high accuracy. SMAP adaptation presents a good compromise between memory need and accuracy.

Table 2 shows the real-time factors measured as actual decoding time[2]. Since the code is not highly optimized with respect to memory and speed this yields only a rough estimate. Due to the increased number of distance calculations, the 8k model is about 50% slower. Note that this difference will decrease if more efficient distance calculation techniques are applied. The increase in CPU demand with the number of Eigenvoices is moderate, adaptation with 8 Eigenvoices is still comparable to MLLR.

4k	8k	MAP	SMAP	MLLR	eigenvoices	
					4	8
0.63	0.93	0.64	0.64	0.71	0.70	0.74

Table 2: Comparison of real-time behavior.

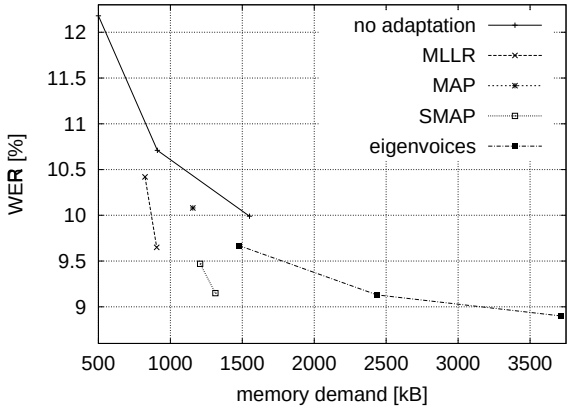

Figure 2: Performance and memory demand for adaptation on 500 words.

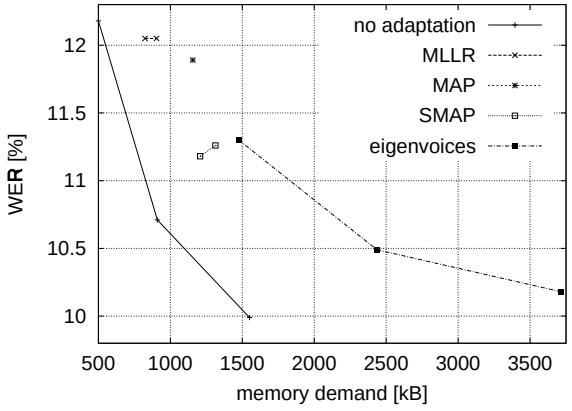

Figure 3: Performance and memory demand for adaptation on 20 words.

5. SUMMARY

We have shown that all investigated adaptation techniques can be applied successfully for adaptation of a compact recognizer with 4k densities. Depending on the amount of adaptation data and memory or processor limitations different adaptation methods can be chosen or a better (bigger) acoustic model may often be used as an alternative to adaptation. If less than 10 words of adaptation data are available, a larger acoustic model outperforms adaptation for all investigated methods.

For 10–40 words of adaptation data the benefits of proper usage of prior knowledge become clearly visible. Eigenvoices perform best for all amounts of adaptation data and achieve by far the best results for less than 50 words. However, they also require the most memory and processing time. SMAP adaptation presents a good compromise between memory need and accuracy.

[2]DIGITAL personal workstation, 500MHz.

6. REFERENCES

[1] Li Lee and R. Rose, "A Frequency Warping Approach to Speaker Normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 49–59, Jan. 1998.

[2] J.W. McDonough and W. Byrne, "Speaker Adaptation with All-Pass Transforms," in *Proc. ICASSP*, 1999, vol. 2, pp. 757–760.

[3] Y. Gao, M. Padmanabhan, and Picheny M., "Speaker Adaptation Based on Pre-Clustering Training Speakers.," in *5th Eurospeech*, 1997, vol. 4, pp. 2091–2094.

[4] X.D. Huang, K.F. Lee, H.W. Hon, and M.Y. Hwang, "Improved Acoustic Modeling with the SPHINX Speech Recognition System," in *Proc. ICASSP*, 1991, vol. 1, pp. 345–348.

[5] K. Yamaguchi, H. Singer, S. Matsunaga, and S. Sagayama, "Speaker-Consistent Parsing for Speaker-Independent Continuous Speech Recognition.," in *Proc. ICSLP*, 1994, vol. 2, pp. 791–794.

[6] J. L. Gauvain and C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, April 1994.

[7] C. J. Leggetter and P. C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[8] V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, W. Byrne, H. Collier, A. Corduneanu, A. Kannan, S. Khudanpur, and A. Sankar, "Rapid Speech Recognizer Adaptation to New Speakers," in *Proc. ICASSP*, 1999, vol. 2, pp. 765–768.

[9] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini, "Eigenvoices for speaker adaptation," in *Proc. ICSLP*, Sydney, 1998, vol. 5, pp. 1771–1774.

[10] H. Botterweck, "Anisotropic MAP defined by Eigenvoices for Large Vocabulary Continuous Speech Recognition," in *Proc. ICASSP*, 2001, vol. 1, p. 353.

[11] K. Shinoda and C.-H. Lee, "A Structural Bayes Approach to Speaker Adaptation," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 276–287, Mar. 2001.

[12] O. Siohan, T. A. Myrvoll, and C.-H. Lee, "Structural Maximum A Posteriori Linear Regression for Fast HMM Adaptation," in *Proc. of ISCA ITRW ASR2000, Automatic Speech Recognition: Challenges for the new Millenium, Paris*, Sept. 2000, pp. 120–127.