# SPEAKER ADAPTATION IN THE PHILIPS SYSTEM FOR LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

*Eric Thelen, Xavier Aubert, Peter Beyerlein*

Philips GmbH Research Laboratories Aachen, Weisshausstrasse 2, 52066 Aachen, Germany
E-mail: {thelen,aubert,beyerlei}@pfa.research.philips.com

## ABSTRACT

The combination of Maximum Likelihood Linear Regression (MLLR) with Maximum *a posteriori* (MAP) adaptation has been investigated for both the enrollment of a new speaker as well as for the asymptotic recognition rate after several hours of dictation.
We show that a least mean square approach to MLLR is quite effective in conjunction with phonetically derived regression classes. Results are presented for both ARPA read-speech test sets and real-life dictation. Significant improvements are reported. While MLLR achieves a faster adaptation rate when only few data is available, MAP has desirable asymptotic properties and the combination of both methods provides the best results. Both incremental and iterative batch modes are studied and compared to the performance of speaker-dependent training.

## 1. INTRODUCTION

Large vocabulary continuous speech dictation systems are likely to be used extensively by just one or a few speakers. Speaker adaptation techniques therefore have to consider the following two issues:

- Speed of short term adaptation with only few data
- Asymptotic performance of long term adaptation with a large amount of user-specific data

Several adaptation methods have been investigated in recent years, belonging either to the Bayesian approach, to transformation techniques or to a combination of both [1,2,3]. Two of them, MLLR [2] and MAP [3] are combined in this paper to address both issues.
When prior distributions are based on the concept of conjugate priors [3], the MAP estimate of a Gaussian mean (under diagonal covariance assumption) is a weighted average of the prior mean and the sample mean. This approach has nice asymptotic properties but the adaptation rate is usually slow as only densities relating to observed acoustic contexts can be updated. On the other hand, the MLLR method uses linear transforms to adjust the initial mean vectors to a new speaker [2]. Depending on the amount of adaptation data, the number of transforms increases from one to several tens or more. This tying of each transformation across a number of densities makes it possible to adapt also distributions for which there were no observations. Hence, all models can be adapted and the adaptation process is dynamically refined when more speaker-dependent material becomes available.
In this paper we show that a least mean square (LMS) approach to MLLR is quite effective in our acoustic modeling as demonstrated by results obtained for ARPA read-speech sets as well as for real-life dictation data. On a variety of tasks, our best results are obtained using a regression class tree derived from a priori phonetic knowledge. For the enrollment of a new speaker using 3 to 13 minutes of speech, we observe a speaker-specific reduction of the error rate in the range of 40-65%. For incremental unsupervised adaptation, the average reduction of the error rate is in the range of 7-15% after a few minutes of speech. For the online long term experiments using real-life data, combined MLLR and MAP adaptation reduces the error rate by 60%. Iterative batch adaptation leads to further improvements, resulting in accuracies comparable to speaker-dependent training.
The MLLR-LMS algorithm is first briefly described together with our way of combining MLLR and MAP. Second, evaluation results using different setups of MLLR are given for several test-sets of the ARPA '94 data. Third, results are presented for real-life dictation and compared to the performance of speaker-dependent training.

## 2. ALGORITHM

The MLLR (Maximum Likelihood Linear Regression, [2]) technique can be applied (among others) to transform each mean vector of continuous density HMMs with a general affine transformation, i.e. a full matrix

and an offset vector:

$$\mu_{new} = \mathbf{A} * \mu_{old} + \mathbf{b} \quad (1)$$

According to the concept of regression classes, a specific transformation might be estimated separately for subsets of similar mean vectors, thus resulting in a piecewise linear transformation of the model space.
In the Philips system [4], the acoustic modeling is based on mixtures of densities trained under Viterbi criterion and on using a globally pooled variance. Hence, the MLLR algorithm simplifies to a least mean square (LMS) approach [2] and the estimation of one transformation matrix is straightforward as described by the following MLLR-LMS equation:

$$\mathbf{T}(r) = [\mathbf{b}, \mathbf{A}](r) = (\sum_{k=1}^{K(r)} \mathbf{o}_k * \tilde{\mu}_k^T) * (\sum_{k=1}^{K(r)} \tilde{\mu}_k * \tilde{\mu}_k^T)^{-1} \quad (2)$$

In this equation, the sum $(\sum_{k=1}^{K(r)})$ is taken over the observation vectors $\mathbf{o}_k$ and augmented mean vectors $\tilde{\mu}_k = [1, \mu^T]^T$ belonging to the regression class $r$. The correspondence between observation and mean vectors is obtained by Viterbi alignment (using the maximum approximation within each mixture). Full transformation matrices are computed from a LDA feature stream instead of block-diagonal matrices [5].
When combining MLLR and MAP adaptation, we compute the MAP estimates for the density means [6] after the MLLR transformation has been carried out. Hence, for densities that have not been observed while processing the adaptation material only MLLR adaptation is applied according to the regression class the density belongs to. Observed densities are moved towards their maximum likelihood estimate $\mu_{obs}$ derived from the adaptation observations (N: Number of adaptation observations, $\alpha$: MAP adaptation parameter):

$$\mu_{new} = \frac{N}{N+\alpha} * \mu_{obs} + \frac{\alpha}{N+\alpha} * \mathbf{T}(r) * \tilde{\mu}_{old} \quad (3)$$

$$\mu_{obs} = \frac{1}{N} \sum_{i=1}^{N} o_i \quad (4)$$

## 3. EVALUATION ON ARPA'94 DATA

The MLLR-LMS algorithm has been applied to various ARPA test-sets for which there are recent results. Gender-dependent models have been trained on 284 speakers of WSJ0+1 database. We use tied-state modeling of within-word triphones [7] with approx. 5,500 HMM states and 50 Laplacian densities per mixture, resulting in about 10 million acoustic parameters per gender. Unseen triphones are handled by the generalized bottom-up state-tying method described in [7].

The so-called Spokes 0, 3 and 4 [8] have been processed following the prescribed boundary conditions: 5k closed-vocabulary and official trigram language model. Spoke 0 was used only for evaluating our recognition system without adaptation (425 sentences, 20 speakers, trigram perplexity of 70). Our word error rate of 6.1% is comparable to the results of HTK (5.7%) and BBN (6.4%) systems, as reported in [8].
Adaptation experiments have then been carried out for Spokes 3 and 4 concerned respectively with non-native and American native speakers. In case of multiple transformations, the regression class tree has been constructed either from a priori phonetic knowledge (in terms of broad phonetic classes, e.g. BPC-43) or through data-driven clustering of the mean vectors (e.g. VQ-64). The regression classes are dynamically estimated based on a minimum number of observations by traversing the tree bottom-up, from leaves to root. In the tables, the maximum number of regression classes is specified by the figure given after 'MLLR'. MAP adaptation was not applied in our first Spoke 3 tests to focus on MLLR.

Table 1: Spoke 3, supervised adaptation, fast enrollment data (500 words, 3-4' speech), Word Error Rate (WER [%]), Relative Improvement ($\Delta$WER [%])

| S3 (non-native) | no adapt. | MLLR 1 global | MLLR 64 tree VQ | MLLR 43 tree BPC |
|---|---|---|---|---|
| WER % | 22.8 | 19.3 | 16.3 | 13.5 |
| $\Delta$WER % | - | -15 | -29 | -41 |

According to table 1, the best results are obtained using a phonetic regression class tree and we observe a significant gain when going from one class to about 20 (effective) classes. So far clustering-driven regression classes led to inferior results. The improvement of 41% although appreciable is smaller compared to the figure of about 50% achieved by other systems [8]. Pursuing on Spoke 3, MAP adaptation has been applied either alone or in combination with MLLR for two adaptation sets. In the second case, the 40 fast enrollment sentences are augmented with 160 WSJ sentences providing a total of 20' speech per speaker.

Table 2: Spoke 3, supervised (non-native) adaptation, MLLR + MAP for 2 adaptation sets, WER [%]

| adapt. speech | no adapt. | MAP only | MLLR only | MLLR +MAP | $\Delta$WER [%] |
|---|---|---|---|---|---|
| 3-4' | 22.8 | 16.4 | 13.5 | 12.4 | -46 |
| 15-20' | 22.8 | 10.0 | 9.4 | 7.5 | -67 |

It is clear from table 2 that combining MLLR with

MAP is always beneficial especially when more adaptation speech is available. Compared to the no-adaptation score of 22.8%, we observe an overall reduction by a factor of 3, yielding almost the same performances as for US-speakers.
In another series of experiments, we considered unsupervised incremental adaptation of native speakers for different tasks. Our Spoke 4 results (see table 3 below) show comparable improvements with those reported for the HTK system (SI-WER: 7.7%, with adaptation: 6.7%, ΔWER: -13%) [8]. Regarding the type of regression class tree, similar conclusions might be drawn as for Spoke 3.

Table 3: Spoke 4, unsupervised incr. adaptation, US-speakers, 100 utterances (1,750 words, 9-14' speech)

| S4 (native) | no adapt. | MLLR 1 global | MLLR 128 tree VQ | MLLR 43 tree BPC |
|---|---|---|---|---|
| WER [%] | 8.50 | 7.55 | 7.43 | 7.20 |
| ΔWER [%] | - | -11.2 | -12.6 | -15.3 |

Last, we evaluated the adaptation technique on North American Business (NAB) data for unlimited vocabulary with a 64K trigram LM [4]. Unsupervised incremental adaptation has been performed after every sentence, each speaker having spoken about 15 sentences. MAP is combined with phonetically derived MLLR transforms. Note that these H0-P0 results (cf. table 4) have been scored with our internal software that does not take account of allowed splits and merges.

Table 4: NAB94 (H0-P0), unsupervised incremental adaptation, 64k-Trigram, development and evaluation

| NAB94 H0-P0 | Development | | Evaluation | |
|---|---|---|---|---|
| | WER | ΔWER | WER | ΔWER |
| no adaptation | 10.80 | - | 10.23 | - |
| MLLR + MAP | 9.63 | -10.8 | 9.46 | -7.5 |

The relative improvements obtained on NAB corpus for incremental unsupervised adaptation are again comparable to available state-of-the-art results.

## 4. APPLICATION TO REAL-LIFE DICTATION (SHORT & LONG TERM ADAPTATION)

Speaker adaptation has been further applied to real-life dictation starting from speaker-independent acoustic models that were trained on the WSJ0 database (84 speakers). Our results are obtained with real-life dictations from a legal application spoken by a female US-English speaker. There is an additional mismatch due to differences of acoustic channels and microphones between WSJ0 training (Sennheiser close-talking) and real-life conditions (hand-held microphone).
We compare MLLR with MAP alone and a combination of both. The test data of 32 minutes of speech (5,126 words) is independent of the speech used for adaptation. Without adaptation, we observe a speaker-independent word error rate of 29.4%. The adaptation is carried out on-line as described in [6].

Table 5: Supervised adaptation with 2 to 13 minutes of adaptation speech (fast enrollment, WER [%])

| adaptation speech | 2' | 4' | 13' |
|---|---|---|---|
| MAP only | 35.6 | 29.7 | 23.5 |
| MLLR 1 (global) | 20.7 | 20.0 | 19.7 |
| MLLR 1 (global) + MAP | 19.9 | 16.7 | 14.8 |
| MLLR 43 (tree, BPC) | 20.0 | 19.2 | 18.2 |
| MLLR 43 (tree, BPC) + MAP | 19.6 | 16.4 | 14.6 |

The benefit of MLLR is especially significant during the first minutes of the adaptation process. MAP adaptation has a negative effect during the first minutes. However, after two minutes of speech, combining both methods appears to be clearly beneficial. With only 4 minutes of speaker-specific data, the error rate is already reduced by 44%.
While multiple MLLR regression classes are superior if only MLLR is used, the combination with MAP adaptation works already very well with only one global regression class for short term adaptation.
We further carried out long term experiments using the same speaker and a maximum of 256 minutes of adaptation speech. Table 6 gives the results for those experiments compared to speaker-dependent training on the adaptation material (SD). As speaker-dependent training uses the knowledge about the spoken text of the adaptation material, most of our experiments have been carried out supervised (SUP). For comparison, we also present the result of unsupervised combined MLLR and MAP adaptation in table 6 (UNSUP).

Table 6: Long-term adaptation (WER [%])

| adapt. speech | SD | MAP only | MLLR only | MLLR + MAP | MLLR + MAP |
|---|---|---|---|---|---|
| | SUP | SUP | SUP | SUP | UNSUP |
| 29' | 23.0 | 18.8 | 16.3 | 13.2 | 16.2 |
| 116' | 11.7 | 14.0 | 15.6 | 11.3 | 14.2 |
| 256' | 9.7 | 12.8 | 15.7 | 11.4 | 13.1 |

MAP adaptation is important for the long term performance, but with the combined method we achieve the best result (error rate reduction 61%). It seems however, that the supervised combined MLLR and MAP

adaptation process is already saturated after two hours of adaptation speech while the unsupervised procedure still takes advantage of additional material to enhance the models.
Unsupervised adaptation leads to inferior results compared to supervised adaptation, but the adaptation process works well despite the recognition errors (WER 7.6%) made on the adaptation material.
Speaker-dependent training outperforms the adapted models when a large amount of speaker-specific material is available. Actually, even with many hours of adaptation speech there are two major differences between adaptation and speaker-dependent training.
The first one is the higher acoustic resolution achieved by a speaker-dependent (SD) training based on a data-driven density splitting technique. Indeed, less than 70% of all densities are observed during adaptation.
The second difference is that during training we iteratively process the data several times. Each iteration allows an improved time alignment between observation vectors, HMM-states and mixture components. In order to determine the influence of this iterative process on the gap between the performance of SD-training and adaptation, we carried out five steps of batch adaptation starting with the references obtained from the on-line experiments (step '0'). First, we performed a time alignment using the adapted references and the spoken text (or the recognized text in the case of the unsupervised experiment). Then, we iteratively performed batch speaker adaptation using the path obtained with this alignment (table 7).

Table 7: Iterative batch adaptation (WER [%])

| | SD | MAP only | MLLR only | MLLR + MAP | MLLR +MAP |
|---|---|---|---|---|---|
| | SUP | SUP | SUP | SUP | UNSUP |
| 0 | 9.7 | 12.8 | 15.7 | 11.4 | 13.1 |
| 1 | - | 11.5 | 15.5 | 10.6 | 12.8 |
| 2 | - | 11.5 | 15.1 | 10.5 | 12.8 |
| 3 | - | 11.5 | 15.1 | 10.3 | 12.6 |
| 4 | - | 11.6 | 14.9 | 10.2 | 12.7 |
| 5 | - | 11.5 | 14.9 | 10.1 | 12.8 |

After five iteration steps the best adaptation result is very close to the result obtained with speaker-dependent training. The largest improvement is achieved with the first iteration step. Especially the error rates in the tests with MAP adaptation decrease significantly, because supervised MAP batch adaptation is very similar to a reestimation step in the training procedure. Unsupervised batch adaptation is not very effective, because it reinforces the recognition errors.

## 5. CONCLUSIONS

We found that combining MLLR and MAP adaptation is beneficial as soon as there are a few minutes adaptation data to get reasonable maximum likelihood estimates of the means used for MAP adaptation. Regarding MLLR, we achieved our best results using a MLLR regression class tree derived from a priori phonetic knowledge.
Although we obtained large error rate reductions with the adaptation procedures, we were not able to outperform speaker-dependent training using the adaptation material with the on-line adapted models. We showed that a major reason for this is the iterative use of the data during the standard training procedure. In our future work, we intend to use this knowledge in order to improve the on-line adaptation procedure.

## 6. REFERENCES

[1] V. Digalakis and L. Neumeyer: "Speaker Adaptation using Combined Transformation and Bayesian Methods", Proceedings ICASSP 1995, Detroit, MI, U.S.A., pp. 680-683
[2] 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
[3] C.-H. Lee, C.-H. Lin, B.-H. Juang: "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models", IEEE Trans. on Signal Processing, Vol. 39, No. 4, April 1991, pp. 806-814
[4] C. Dugast, P. Beyerlein, R. Haeb-Umbach: "Application of Clustering Techniques to Mixture Density Modeling for Continuous Speech Recognition", Proceedings ICASSP 1995, Detroit, MI, U.S.A., pp. 524-527
[5] L. Neumeyer, A. Sankar, V. Digalakis: "A Comparative Study of Speaker Adaptation Techniques", Proceedings EUROSPEECH, Madrid, Spain, September 1995, pp. 1127-1130
[6] E. Thelen: "Long Term on-line Speaker Adaptation for Large Vocabulary Dictation", Proceedings ICSLP 1996, Philadelphia, PA, U.S.A., pp. 2139-2142
[7] X. Aubert, P. Beyerlein, M. Ullrich: "A Bottom-Up Approach for Handling Unseen Triphones in Large Vocabulary Continuous Speech Recognition", Proceedings ICSLP 1996, Philadelphia, PA, U.S.A., pp. 14-17
[8] D.S. Pallett et al.: "Benchmark Tests for the ARPA Spoken Language Program", Proceedings of the Spoken Language Systems Technology Workshop 1995, Austin, Texas, U.S.A., pp. 5-38