DATABASE AND ONLINE ADAPTATION FOR IMPROVED SPEECH RECOGNITION IN CAR ENVIRONMENTS

Alexander Fischer and Volker Stahl

Philips Research Laboratories, Weisshausstr. 2, D-52066 Aachen, Germany
email: {afischer,vstahl}@pfa.research.philips.com

ABSTRACT

Data collections in the car environment require much more effort in terms of cost and time as compared to the telephone or the office environment. Therefore we apply supervised database adaptation from the telephone environment to the car environment to allow quick setup of car environment recognizers. Further reduction of word error rate is obtained by unsupervised online adaptation during recognition. We investigate the common techniques MLLR and MAP for that purpose. We give results on command word recognition in the car environment for all combinations of database and online adaptation in task-dependent and task-independent scenarios. The possibility of setting up speech recognizers for the car environment based on telephone data and a limited amount of adaptation material from the car environment is demonstrated.

1. INTRODUCTION

Automatic speech recognition offers increased safety and user comfort through hands-free and eyes-free operation of the car functionality and communication/entertainment equipment. However, despite remarkable progresses in the past years, environmental noise is still one of the most challenging problems for reliable speech recognition. As is the case with many real life problems there are two ways to handle them: getting rid of them or trying to adapt. In the literature we find these two approaches (and combinations thereof) applied to the problem at hand: reduce the noise during the feature extraction (e.g. spectral subtraction [1]) or adapt the acoustic models (usually HMM's) to the noisy environment. In this paper we concentrate on the second approach. A trivial way to obtain environment adapted HMM's is to use training data collected in the target environment. While this leads to minimal error rates, the cost for the data collection is in many cases prohibitive. Further, it is not always possible to predict the precise noise conditions where the recognizer will be used, i.e. it is important that the system is able to adapt to its environment while being in operation. Much research has therefore been devoted on how to modify the parameters of a HMM which was trained in a clean environment such that it works well under noise. Good results have been achieved by a technique called parallel model combination (PMC) see e.g. [2, 3, 4, 5, 6, 7]. However to avoid training of separate noise models and the computational complexity in PMC, in this paper we use the well known MLLR [8] and MAP [9, 10] algorithm to adapt HMM's trained with telephone material to the car environment. We consider both supervised adaptation (in the following called *database adaptation* using a small database of speech material collected in different cars and by different speakers as well as unsupervised adaptation (in the following called *online adaptation*) to a concrete car and speaker while the recognizer is in use. As database adaptation requires the collection of car speech material we investigate how much adaptation material is required in order to obtain comparable results to training exclusively on pure car material. Further, we have to distinguish whether the adaptation material and the test set are from the same application domains (*task dependent adaptation*) or not (*task independent adaptation*). Task independent adaptation leads usually to higher error rates than task dependent adaptation but is the method of choice when new speech applications are developed for which no specific speech data is available. Similar experiments were conducted e.g. in [7] using PMC, however instead of real car speech data a synthetic addition of car noise and clean speech was used, which is only an approximation to the real environment.

In our experiments we employ the Philips continuous speech recognition system as described in [13] with a nonlinear spectral subtraction enhanced front-end [11, 12]. Maximum likelihood linear regression [8] proves to be very powerful for both database and online adaptation due to its generalization capabilities. MAP [14] seems to be more sensitive to the choice of parameters than MLLR. Best results are obtained by a proper combination of both.

Section 2 briefly reviews the adaptation framework for MLLR and MAP employed for database and online adaptation. The experimental setup and the utilized databases are described in Section 3. After presentation of the experimental results in Section 4 and their discussion we finish with a short conclusion and future perspective in Section 5.

2. REVIEW OF MLLR AND MAP ADAPTATION

This section contains a brief review of the MLLR (maximum likelihood linear regression) and MAP (maximum a posteriori) adaptation method. In our experiments we use a simplified version of MLLR and MAP in the sense that only the mean vectors of the HMM emission distributions are adapted but not the covariances or other parameters and only a single MLLR regression class is used [15]. For a more general definition of the methods see e.g. [8, 9, 16, 10, 14, 17].

MLLR. A MLLR adaptation step consists of the estimation of a linear affine transform A, b and its application to all emission distribution means μ:

$$\mu_{\mathrm{new}} = A\mu_{\mathrm{old}} + b.$$

The linear transform is given by

$$[b, A] = \left(\sum_{i=1}^{N} o_i \tilde{\mu}_i^T \right) \left(\sum_{i=1}^{N} \tilde{\mu}_i \tilde{\mu}_i^T \right)^{-1}$$

where N is the number of observation vectors o_i and corresponding augmented mean vectors $\tilde{\mu}_i = [1, \mu_i^T]^T$. This transform is optimal in the sense that it minimizes

$$\sum_{i=1}^{N} ||o_i - \mu_{\mathrm{new}_i}||^2.$$

MAP. A MAP adapted mean vector is a weighted average of the prior mean and the mean of the adaptation observations:

$$\mu_{\mathrm{new}} = \frac{N\alpha}{N\alpha+1}\mu_{\mathrm{obs}} + \frac{1}{N\alpha+1}\mu_{\mathrm{old}} \quad (1)$$

$$\mu_{\mathrm{obs}} = \frac{1}{N}\sum_{i=1}^{N} o_i. \quad (2)$$

The parameter α defines the "adaptation speed", i.e. the weight of new observations μ_{obs} as compared to the old estimation μ_{old}.

In both methods the correspondence between observation and mean vectors is obtained by Viterbi alignment.

MLLR + MAP. When comparing MLLR and MAP one often finds that MLLR works well already for few observations whereas MAP is asymptotically better. This can be explained by the global transform of MLLR, which results in an adaptation of all mean vectors, even if few or no observations of a particular mean vector are available. On the other hand, as a global transform is a rather coarse approach, MAP is more accurate in the presence of many observations. In the experiments reported below the following combination of MLLR and MAP is applied:

$$\mu_{\mathrm{new}} = \frac{N\alpha}{N\alpha+1}\mu_{\mathrm{obs}} + \frac{1}{N\alpha+1}(A\mu_{\mathrm{old}} + b).$$

A MLLR+MAP adaptation step is carried out at the end of each utterance. However, as the reliable estimation of an MLLR transform requires a certain amount of observations, a parameter N_{obs} is introduced and the first MLLR adaptation is applied only after N_{obs} observations.

The same adaptation algorithm is used for database and online adaptation. The essential difference between database and online adaptation is that for database adaptation the correct transcription is known (supervised adaptation) whereas online adaptation relies on the recognition result (unsupervised adaptation) and therefore it can happen that a wrong transcription is used for online adaptation.

3. EXPERIMENTS

We utilize the SpeechDat [18] database for the telephone environment and the MoTiV Car Speech Data Collections (CSDC) database [19] for the car environment. MoTiV [20] is a project funded by the German Federal Ministry of Education, Science, Research and Technology focusing on mobility and transportation in intermodal traffic systems. Automatic speech recognition as part of a user-friendly human-machine interface is one of the subprojects. Our interest in this investigation is to study how telephone databases can be used as starting point for the development of automatic speech recognizers for the car environment. We will investigate the following scenarios:

- Task- and environment dependent adaptation in the car (Table 2 upper block)
- Task-independent and environment dependent adaptation in the car (Table 2 middle block)
- Task-independent and environment in-dependent adaptation (telephone → car, Table 2 lower block)

In all three cases we will combine database adaptation with online adaptation. Online adaptation in our experiments is unsupervised adaptation of the references at the end of each sentence with a combination of MLLR with one regression class and MAP. Preliminary investigations have shown that the MAP movement factor α has the main influence on the adaptation performance. The settings of MLLR (one regression class and $N_{obs} = 400$) could be fixed and were found to be quite robust. For database adaptation MLLR is always applied. In online adaptation all combinations (also excluding one of the methods) are investigated.

Training of the telephone recognizers is done on the SpeechDat phonetically rich sentences combined with command word phrases (tel phon, 600 speakers, 11 sentences each). Car environment references are trained on isolated command words (car cmd, 116 speakers in 3 cars, 43 utterances each, 43 words in total) or phonetically rich sentences (car phon, 205 speakers in 3 different cars, 9 sentences each).

Database adaptation sets are the car command words training set (cmd) for task-dependent adaptation and a "generic" vocabulary set of various isolated words (city names, commands, names, etc.) in the car (gen) is used for task-independent adaptation. The car command words test set (car cmd, 39 speakers, 43 utterances each, 43 words in total) is the same in all experiments.

4. RESULTS

First we give some baseline results for the telephone and car environment [12] without any adaptation techniques (upper block of Table 1). Each line in Tables 1 to 2 states the training data, type of acoustic models (*word* for whole word models or *phon* for monophones), database adaptation data, test data, online adaptation settings (- means none, 0.0 implies MLLR only and nonzero stands for a combination of MLLR and MAP with the specified movement factor) and the obtained WER. In the lower block of the table online adaptation is applied to the upper block scenarios.

train	model type	dat. adapt	test	α_{OA}	WER [%]
car cmd	word	-	car cmd	-	2.76
car cmd	phon	-	car cmd	-	5.78
car phon	phon	-	car cmd	-	14.46
tel phon	phon	-	car cmd	-	21.74
car cmd	word	-	car cmd	0.0	1.80
car cmd	phon	-	car cmd	0.0	4.88
car phon	phon	-	car cmd	0.0	10.48
tel phon	phon	-	car cmd	0.4	11.18

Table 1: Word error rates without database adaptation (α_{OA} = 0.0 means MLLR only)

We can see significant improvement for whole word and phoneme models in the car environment (15%-35%) relative). The preferable online adaptation technique in these scenarios is MLLR without MAP. Direct usage of telephone environment monophones yields a WER of 21.74% that can be reduced by almost 50% relative to 11.18% using a combination of MLLR and MAP with α_{OA}=0.4. The resulting WER is only marginally above the car environment trained phoneme references. Note, that in the last

two lines we obtain almost the same word error rates and we did not use any car environment training data in the last experiment!

Further improvement in the task-independent case (last two rows of Table 1) can be achieved through database adaptation. We distinguish two cases: 1) We sacrifice the task-independence of initial recognizer assuming the availability of a data collection of the target vocabulary (car cmd training set in our case) in the car environment in order to improve recognition performance. 2) If we want to keep task-independence we have to use generic adaptation material (gen). An important question is the necessary amount of the adaptation material in the car environment (i.e. number of speakers = cost).

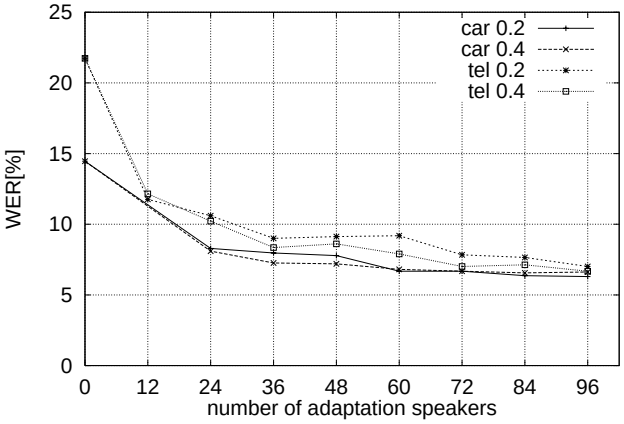

Figure 1: Task-dependent adaptation of car and telephone monophone references without online adaptation. The starting point (0 adaptation speakers) is given by the results in Table 1. The numbers behind the environment represent the MAP adaptation factor α.

Figures 1 and 2 show the performance improvement in WER over the number of adaptation speakers (data collection size) both for car and telephone environment phoneme references without online adaptation as starting point. Figure 1 covers the task-dependent case (adaptation on car cmd corpus) and Figure 2 contains the task-independent case (adaptation on car gen corpus).

As expected, we observe continuous performance improvement with increasing number of adaptation speakers in the task-dependent case (Figure 1). We finally obtain WERs similar to task-dependent training. So task-dependent adaptation converges to task-dependent training when the size of the adaptation material approaches the size of task-dependent training material. The parameter settings in this type of database adaptation are not crucial. We see only minor variation over the MAP movement factors $\alpha_{\text{DA}} = 0.2, 0.4$. The performance is similar for $\alpha_{\text{DA}} = 0.6, 0.8$ which is not shown in the figure. In the task-independent scenarios (Figure 2) $\alpha_{\text{DA}} = 0.2$ turns out to be the best setting both for the telephone and the car environment as starting point. The curves for $\alpha_{\text{DA}} > 0.4$ reveal significantly worse performance and are not shown for sake of clarity. For small amounts of adaptation material (less than 20 speakers) performance degradation occurs. Beyond 24 speakers slow but steady improvement takes place. With 40-60 speakers of adaptation material (being a reasonable size in terms of collection effort) 50%-90% (car) respectively 78%-85% (tel) of the total performance gain (with 205 adaptation speakers) is obtained. The initial degradation for few adaptation speakers can be explained by the primary focus on speakers in this phase of database adaptation which then gradually moves to environment adaptation for larger numbers of speakers. Of course, the above pure database adaptation scenarios can be combined with online adaptation. Table 2 summarizes the major results of such combination scenarios for some numbers of (database) adaptation speakers (24, 48, 72).

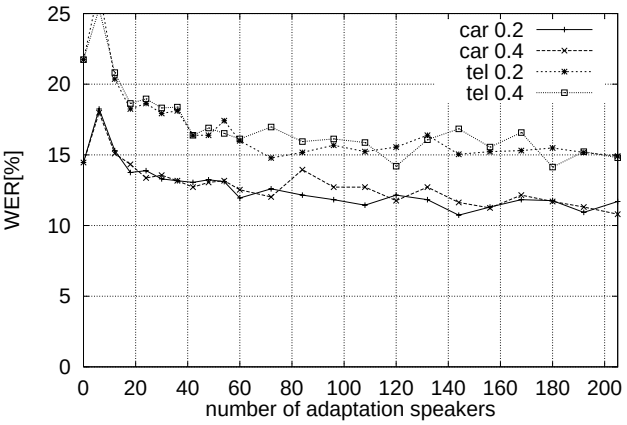

Figure 2: Task-independent adaptation of car and telephone monophone references without online adaptation (figures = MAP alpha). The starting point (0 adaptation speakers) is given by the results in Table 1. The numbers behind the environment represent the MAP adaptation factor α.

For task-dependent database adaptation (cmd) in the car environment the major gain by online adaptation is due to MLLR (approximately two thirds of the overall gain by online adaptation). The total improvement of 15%-20% relative is obtained by combined MLLR and MAP with $\alpha_{\text{OA}} = 0.2$. In the task-independent database adaptation case (gen) the above relation turns around. Here MLLR in online adaptation contributes to approximately one third and additive MAP with $\alpha_{\text{OA}} = 0.2$ as optimal value plays the dominant role. The overall performance improvement is about 35%-45% relative. Furthermore a minimum number of speakers in the adaptation material is necessary to achieve positive effects. If environment adaptation (tel $\rightarrow$ car) is also part of the database adaptation process, the latter effect is even strengthened. MLLR online adaptation yields 2%-3% relative improvement whereas the combination with MAP with $\alpha_{\text{OA}} = 0.2$ leads to the overall improvement of 30%-40% relative.

5. CONCLUSION

We have demonstrated how the adaptation techniques MLLR and MAP can be used to significantly improve automatic speech recognition in a car environment. Depending on the available training and adaptation material we defined two adaptation strategies: database and online adaptation. We quantified the expectable performance gains by each of those strategies and their combination on an isolated command word recognition task. The performance of task-independent car environment references was improved from 14.46% word error rate to 6.43% (55.5% relative) through task-dependent database adaptation (with 48 speakers adaptation material) combined with online adaptation. In order to avoid costly

train	model type	dat. adapt	test	α_{OA}	WER [%]
car phon	phon	cmd24	car cmd	-	8.87
car phon	phon	cmd24	car cmd	0.0	7.39
car phon	phon	cmd24	car cmd	0.2	7.01
car phon	phon	cmd48	car cmd	-	7.58
car phon	phon	cmd48	car cmd	0.0	6.68
car phon	phon	cmd48	car cmd	0.2	6.43
car phon	phon	cmd72	car cmd	-	6.56
car phon	phon	cmd72	car cmd	0.0	5.91
car phon	phon	cmd72	car cmd	0.2	5.59
car phon	phon	gen24	car cmd	-	13.88
car phon	phon	gen24	car cmd	0.0	12.08
car phon	phon	gen24	car cmd	0.2	7.71
car phon	phon	gen48	car cmd	-	13.24
car phon	phon	gen48	car cmd	0.0	11.95
car phon	phon	gen48	car cmd	0.2	8.68
car phon	phon	gen72	car cmd	-	12.60
car phon	phon	gen72	car cmd	0.0	10.22
car phon	phon	gen72	car cmd	0.2	8.23
tel phon	phon	gen24	car cmd	-	18.64
tel phon	phon	gen24	car cmd	0.0	18.06
tel phon	phon	gen24	car cmd	0.4	10.86
tel phon	phon	gen48	car cmd	-	16.39
tel phon	phon	gen48	car cmd	0.0	16.00
tel phon	phon	gen48	car cmd	0.4	10.54
tel phon	phon	gen72	car cmd	-	14.78
tel phon	phon	gen72	car cmd	0.0	16.26
tel phon	phon	gen72	car cmd	0.2	9.45

Table 2: Word error rates with database adaptation (α_{DA} = 0.2) and optional online adaptation (α_{OA} = 0.0 means MLLR only). The upper block relates to Figure 1 and the middle and lower block relate to Figure 2.

car environment data collection we investigated the use of (available) telephone environment data as starting point. The task-independent telephone environment references could be improved from 21.75% word error rate to 10.86% (50.1% relative) by combined task-independent database adaptation and online adaptation. Further improvement can be expected by the use of office instead of telephone environment data as starting point and the use of supervision information in online adaptation through confidence measures or explicit feedback from the user-interface.

6. REFERENCES

[1] A. N. Flores and S. Young, "Continuous Speech Recognition in noise using spectral subtraction and HMM Adaptation," in *Proc. ICASSP*, pp. 409–412, 1994.

[2] M. Gales and S. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, vol. 12, pp. 231–239, 1993.

[3] M. Gales and S. Young, "HMM recognition in noise using parallel model combination," in *Proc. EUROSPEECH*, pp. 837–840, 1993.

[4] M. Gales and S. Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," *Computer Speech and Language*, vol. 9, pp. 289–307, 1995.

[5] R. Yang and P. Haavisto, "Noise compensation for speech recognition in car noise environments," in *Proc. ICASSP*, pp. 433–436, 1995.

[6] R. Yang and P. Haavisto, "An improved noise compensation algorithm for speech recognition in noise," in *Proc. ICASSP*, pp. 49–52, 1996.

[7] L. Neumeyer and M. Weintraub, "Robust speech recognition in noise using adaptation and mapping techniques," in *Proc. ICASSP*, pp. 141–144, 1995.

[8] 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.

[9] 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, April 1994.

[10] C. H. Lee, C. H. Lin, and B. H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Transactions on Signal Processing*, pp. 806–814, Apr. 1991.

[11] D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach, and T. Eisele, "Investigation of acoustic front ends for speaker-independent speech recognition in the car," in *Proc. EUROSPEECH*, pp. 2571–2574, 1997.

[12] A. Fischer and V. Stahl, "Subword unit based speech recognition in car environments," in *Proc. ICASSP*, pp. 257–260, 1998.

[13] V. Steinbiss, H. Ney, X. Aubert, S. Besling, C. Dugast, U. Essen, D. Geller, R. Haeb-Umbach, R. Kneser, G. Meier, M. Oerder, and B.-H. Tran, "The Philips research system for continuous–speech dictation," in *Philips Journal of Research*, pp. 317–352, 1995.

[14] C. H. Lee and J. L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," in *Proc. ICASSP*, pp. 558–561, 1993.

[15] E. Thelen, "Long term on-line speaker adaptation for large vocabulary dictation," in *Proc. ICSLP*, pp. 2139–2142, 1996.

[16] L. Neumeyer, A. Sankar, and V. Digalakis, "A Comparative Study of Speaker Adaptation Techniques," in *Proc. EUROSPEECH*, pp. 1127–1130, 1995.

[17] Q. Huo and C. Chan, "On-line Bayes adaptation of schmm parameters for speech recognition," in *Proc. ICASSP*, pp. 708–711, 1995.

[18] H. Hoege, H. Tropf, R. Winsky, H. van der Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," in *Proc. ICASSP*, pp. 1771–1774, 1997.

[19] D. Langmann, T. Schneider, R. Grudszus, A. Fischer, T. Crull, H. Pfitzinger, M. Westphal, and U. Jekosch, "CSDC - The MoTiV Car-Speech Data Collection," in *First International Conference on Language Resources and Evaluation*, 1998.

[20] TÜV Rheinland Sicherheit und Umweltschutz GmbH, "MoTiV - Mobilität und Transport im intermodalen Verkehr." http://www.tuev-rheinland.de/tsu/bvt/ motiv/haupts.htm.