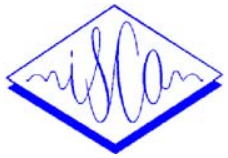
ISCA Archive
http://www.isca-speech.org/archive



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DOMAIN ADAPTATION FOR ROBUST AUTOMATIC SPEECH RECOGNITION IN CAR ENVIRONMENTS

Rolf Bippus, Alexander Fischer, Volker Stahl

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

10.21437/Eurospeech.1999-427

ABSTRACT

A major obstacle for the migration of automatic speech recognition into every-day life products is environmental robustness. Automatic speech recognition systems work reasonably well under clean (laboratory) conditions but degrade seriously under real world conditions (e.g. out-door, car). A lot of research work is devoted to increase the environmental robustness of automatic speech recognition systems. A common method is to use clean (office) data as a starting point and simulate the degraded environmental situation by additive artificial (e.g. Gaussian) or recorded noise from the real environment [1]. We study the validity of such additive noise experiments with regard to a real noisy environment. With regard to a previously published work on database adaptation we also examine the possible benefit when using models trained in the simulated environment as a starting point for adaptation ([2]). We present experimental results on data recorded for task-dependent whole word and phoneme modeling in the car environment on data from the the MoTiV Car Speech Data Collection (CSDC) [3].

1. INTRODUCTION

Despite remarkable progresses in the past years, environmental noise is still one of the most challenging problems for reliable automatic speech recognition. Especially the car environment is a challenging domain for automatic speech recognition systems due to its high level of different types of background noise (engine, car body, ventilation).

Approaches for handling this problem mainly fall into two categories: Either trying to get rid of the noise beforehand or adapting the recognizer to the actual noisy environment. Both approaches are found in the literature using e.g. spectral subtraction, adaptive filtering in the feature extraction stage or adapting the acoustic models by various methods. Concentrating on the latter, good results have been achieved for example by Parallel Model Combination (PMC) ([4]), but also standard adaptation techniques like MLLR and MAP adaptation may be used successfully ([5], [6], [2]).

In all adaptation scenarios however the question arises how to obtain models that may best serve as a starting point for online adaptation. Especially for MLLR, MAP and lately proposed extensions to PMC ([7]) this is a crucial point. Models trained on data actually collected within the targeted environment might best serve the purpose but due to the high costs that might come with such a data collection or in cases where the target environment might not be known or changing, this may not be feasible. Considering this our main interest in this study is targeted at the following question: How well can speech recognizers trained on clean data with additive (artificial or real) noise (we will call this the simulated environment) perform in the real environment. As far as this study is concerned, the following additional constraints apply to the resulting recognizer: It should be speaker independent and environmental independent with respect to the set of different environments it might be supposed to operate in (in our case, we know this will be in a car, but it should be as independent as possible of the specific car or driving situations).

Our investigation starts with models trained on clean data from an office environment. Their performance is tested on data from the real car environment at a great variety of driving situations and SNR levels, giving an lower bound for the performance. In contrast a somewhat upper bound for the performance is obtained by using models that are trained on data recorded under identical conditions as the testing data.

Then the effect of simulating the real environment by adding artificial or real noise to the clean office data and using this for training is investigated. Assuming that it might be feasible to collect at least a small amount of training data from the real target environment, in a last step the best models obtained from the simulated environment are used as a starting point for supervised MLLR and MAP adaptation (database adaptation, [2]). This is done in order to examine the benefit when using models that already represent the target environment to some extent (by being trained in the simulated one) as a starting point over starting with clean models.

Section 2 gives the details on the experimental setup. Section 3 presents the major results of the experiments and section 4 concludes and gives further perspectives.

2. EXPERIMENTAL SETUP

All experiments were carried out using two different, small vocabulary, isolated word recognition tasks for which a sufficient amount of training and test data was available collected in both office and noisy car environment. Thus basically two different datasets were used for each task, one collected in clean office environment serving as training data, and the second collected in the real car environment basically serving as testing data. There is no speaker overlap in the training/testing datasets. In order to

achieve the highest possible significance level of the results obtained, all experiments were performed in such a way, that basically the complete real car environment data was used for testing. In those experiments, where part of this data was needed for training and/or adaptation, this meant to split the data into several parts that in turn were used for testing using the remaining part for training/adaptation and averaging over the different results (leave-n-out method). This makes sure that all WER results are obtained on essentially the same testing data.

For each task, experiments were carried out using whole word models and monophone models, in both cases trained task dependently, that is trained and tested on identical vocabulary. The experiments using the monophone models should give an idea of what loss is to be expected when switching from word models to monophone models, the number of parameters being only approximately 1/3 of that of the word models.

In the experiments presented no kind of online adaptation was used. We used nonlinear spectral subtraction ([8]) and cepstral mean subtraction throughout all experiments.

2.1. SNR Estimation

Due to the fact that our system generally applies a pre-emphasis filter to the signal all given SNR estimates are based on the signal after filtering. This especially has the effect that for artificial, uncorrelated, white noise (like white Gaussian noise), the SNR is about 3 dB lower than it would be on the unfiltered signal since the energy of the noise is doubled by the pre-emphasis filter. On the other hand for real car noise measurements gave an SNR approximately 2 to 3 dB higher than on the unfiltered signal due to the high energy components at low frequencies which are partly suppressed by the high pass characteristic of the filter.

All estimations of the SNR of a speech signal are based on a speech/pause segmentation given by forced alignment of the truly spoken text with the signal (Viterbi alignment), using models trained on the same dataset.

2.2. Noise Addition

For simulating the real environment real car and white Gaussian noise is added to the signals prior to pre-emphasis at a specified target SNR that the combined signal should have after pre-emphasis. Therefore in accordance to the above the weights for noise addition are calculated based on the energies obtained from the pre-emphasized signals.

In the case of real noise, we used a mix of different noise recordings from different cars and driving situations. For each utterance in the database one of these noise recordings was randomly chosen for noise addition.

2.3. Database Adaptation

What follows is 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 [9]. An MLLR adaptation step consists of the estimation of a linear affine transform A, b and its application to all emission distribution means μ:

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

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

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

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

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

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

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:

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

In contrast to online adaptation, for database adaptation the correct transcription is known (supervised adaptation), thus errors due to a wrong transcription used for adaptation are impossible. In the experiments below we use an iterative procedure for database adaptation. An MLLR+MAP adaptation step is carried out after all data in the adaptation dataset has been processed (Viterbi Alignment). Using the newly adapted references this is repeated for a fixed number of iterations.

2.4. Database

The two isolated word recognition tasks consist of 37 German city names (cities37) for the first and 41 possible commands for the control of a car navigation system (commands41) for the second task.

For each task two datasets were used, one containing the data collected in an office environment denoted by cities37_office and commands41_office, respectively, the other one containing the corresponding car environment data, cities37_car and commands41_car.

The data stems from the MoTiV [10] Car Speech Data Collection (CSDC) [3]. MoTiV is a project funded by the German Federal Ministry of Education, Science, Research and Technology. The car data was collected in 3 different cars under a great variety of driving situations, at SNRs of approximately 11 ± 5 dB (mean $\pm$ standard deviation) covering a total range from -4dB to 30dB. The same utterances were collected in an office environment, spoken by different speakers at an average SNR of approximately 21 ± 5 dB. The average mismatch in SNR between office and car data is thus about 10 dB. The following table shows the main facts for the different databases.

	car	office
cities37	5734 utterances 156 speakers	6248 utterances 88 speakers
commands41	5711 utterances 156 speakers	3885 utterances 75 speakers

The real noise that was used for noise addition stems from the same data collection and was recorded from a total of 3 cars in 3 different driving situations (speeds). It is important to notice that none of these cars was present in the testing dataset in order to emphasize the environmental independence of the resulting models.

3. RESULTS

3.1. Baseline

The following table shows the baseline results that serve as an upper respectively lower bound for the remaining experiments. In order to get a feeling for the difficulty of the tasks, the first column (office-office) gives the WER obtained for the matched office scenario, training and testing being performed on the office data. The second column (office-car) gives the lower performance baseline, obtained with models trained on the clean office data and no additional measures taken (except nonlinear spectral subtraction). The last column (car-car) serves as an upper bound for performance when training the models under real environment conditions. These results were obtained by leaving–1/3–out as described above.

\train - test task + models \	office-office	office-car	car-car
orte37 words	0.62	8.77	1.26
app_nav41 words	0.53	7.77	2.57
orte37 monophones	0.83	17.21	3.09
app_nav41 monophones	1.27	14.38	5.99

Table 1: Baseline Results.

3.2. Additive Noise

Here the results on the simulated car environment using additive noise are presented. The first row repeats the baseline results from above, when no measures are taken. The following rows show the comparison between the addition of artificial, Gaussian, white noise (GAU) and mixed real car noise (CAR) as described in section 2 at a number of different target SNRs. The bold faced numbers show the optimal target SNR value for each scenario respectively.

\task / models noise type / SNR \	orte37 words	 phn	app_nav41 words	 phn
none	**8.77**	**17.21**	**7.77**	**14.38**
GAU / 3dB	6.19	14.71	6.61	15.15
GAU / 6dB	**5.77**	**13.77**	5.03	12.31
GAU / 9dB	6.99	13.84	**5.45**	**12.03**
GAU / 12dB	8.87	16.83	7.52	13.54
GAU / 15dB	14.78	19.20	9.56	17.23
CAR / 3dB	7.23	17.01	11.29	20.85
CAR / 6dB	4.56	10.85	5.91	9.46
CAR / 9dB	**4.00**	10.30	**3.97**	**7.52**
CAR / 12dB	5.04	**9.11**	4.61	8.30
CAR / 15dB	7.30	10.99	6.96	9.28

Table 2: Additive Noise results for artificial white Gaussian noise (GAU) and a mix of noises from different cars (3 cars not present in testing data, CAR) added to clean office training data at different target SNRs

The best results were obtained using the real noise mix at a target SNR that approximately corresponds to the average SNR of the testing database. They are about 30% relative better than those using artificial noise. Additional experiments using real noise from a single car not present in the testing database show results somewhere between artificial noise and mixed real noise, depending on the car actually used. The results are not presented in table 3.2 but in all cases are consistently better than using artificial noise and worse than using the real noise mix.

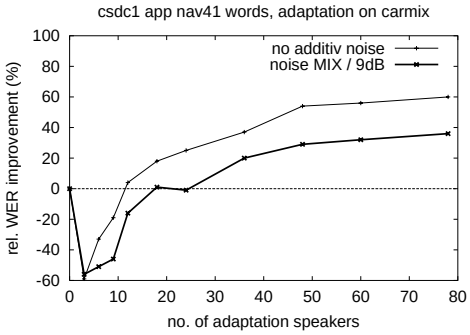
3.3. Database Adaptation

In order to apply database adaptation as described above, the testing databases were now split into 2 parts each and one part was in turn used for supervised MLLR+MAP adaptation the other for testing. In this case, data from the identical 3 cars was present in the adaptation and testing material. However the speakers were distinct for adaptation and testing. Thus in each case the adaptation set contained about half the utterances from 78 speakers. The results in table 3.3 below were obtained for a fixed number of 20 iterations and a MAP adaptation speed of $\alpha = 0.4$. The given noise type corresponds to the one used for noise addition when training the models that were used as the starting point for the adaptation. Results are given for starting from the optimal results achieved by noise addition (see above). The row for full training repeats the results from above when fully training the models in the car environment.

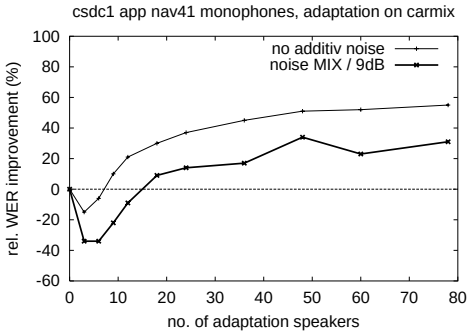
\task / models noise type / SNR \	orte37 words	 phn	app_nav41 words	 phn
no noise	2.67	5.23	3.09	6.50
CAR / 9dB	1.69	-	2.53	5.73
CAR / 12dB	-	4.67	-	-
full training	1.26	3.09	2.57	5.99

The important point to note here is that starting with references trained in the simulated car environment not only gives consistently better results but also that the obtained results are not far from the upper bound established by a full training on car data. In case of the command words even about the same results could be achieved.

In order to answer the question of how much adaptation data is needed in order to achieve substantial enhancements by database adaptation, the following figures show the relative improvement in word error rate versus the number of different speakers in the adaptation set.



(a) app_nav41, word models, carmix adaptation



(b) app_nav41, monophones, carmix adaptation

Figure 1: Relative improvement in WER using database adaptation with different number of speakers. Each plot shows two curves, one for starting from models trained on clean office data (no additive noise), the other for starting from models trained in the simulated environment (noise CAR / 9(12)dB).

It can be seen in all cases that the number of speakers used for adaptation has to be considerably higher in case the simulated training environment was used for training the starting references. (keeping in mind that the figures show only relative improvement and the absolute word error rates considerably differ, see absolute results above). This leads to the conclusion that the environment is already modelled to some extend and the adaptation mainly focuses on the speakers in the adaptation set. A dominating effect of environmental adaptation is observed only for a far larger number of speakers compared to starting with clean references. This effect is not quite as strong when using monophones probably due to the overall smaller number of mixture densities that goes with a certain smoothing effect and only allows for less speaker specific adaptation.

4. CONCLUSION

We have demonstrated that simulating the real environment by additive real car noise is a valid approach for obtaining models that cover an unknown car environment to quite a good extent. It works considerably better compared to the use of artificial noise. Real car noise may be cheaply obtained even for quite a large number of cars and different driving situations.

In cases where additional speech data from the real environment is available additional database adaptation may achieve additional gain provided sufficient adaptation material is available. When starting with models that were trained on the simulated noisy environment about twice as much adaptation material is needed compared to starting with clean references, otherwise adaptation might even result in a considerable degradation in the case of speaker independent recognition. In case sufficient adaptation material is available, training in simulated environment and MLLR+MAP adaptation produces results quite close to the baseline established by full training in the real target environment. In case of the command words we even achieved about equally good results.

The experiments presented will be extended towards using car specific noise for noise corruption and performing car specific adaptation. Further work will be directed to the setup of task-independent recognizers based on the presented work.

5. REFERENCES

[1] Varga, A.; Steeneken, H. J. M.; Tomlinson, M.; Jones, J. ”The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition”, Booklet included in the NOISEX-92 CD-ROM Set.

[2] Fischer, A.; Stahl, V. ”Database and Online Adaptation for Improved Speech Recognition in Car Environments”, Proceedings of ICASSP, Phoenix, USA, Vol. 1, pp. 445 - 449, 1999.

[3] 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.

[4] M. J. F. Gales, “Model-Based Techniques fro Noise Robust Speech Recognition”, *Dissertation, University of Cambridge*, 1995.

[5] 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.

[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, April 1994.

[7] Crafa, S.; Fissore, L.; Vair, C. ”Data-Driven PMC and Bayesian Learning Integration for Fast Model Adaptation in Noisy Conditions.”, Proceedings of ICSLP, Sydney, Australia, Vol. 2, pp. 471-473 ,1998

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

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

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