TOWARDS EXPLOITING THE POTENTIAL OF ENVIRONMENT ADAPTATION

Christian Geißler, Josef G. Bauer

Siemens AG, Corporate Technology
Munich, Germany
{christian.geissler, josef.bauer}@siemens.com

ABSTRACT

The offline HMM adaptation of a generic car speech recognizer to a specific car environment is investigated. For the generation of the adaptation database the approach of Environment Adapted Databases (EADB) is applied that avoids real speech recordings in the target environment and therefore reduces the effort significantly. With MLLR adaptation using such an EADB a relative reduction of the word error rate of more than 10% can be achieved on a British city names task. It is proven by adaptation on real speech recordings from the target environment that the improvement with EADBs fully exploits the potential of HMM adaptation for the given car. Additionally it can be shown that if task matching material is available for adaptation a performance improvement of more than 30% can be reached with an additional maximum likelihood training iteration.

1. INTRODUCTION

Speech recognition applications seem to become more and more popular for the use in vehicles. A particular property of the car environment is that training databases are often recorded in a limited number of car types and with a particular microphone. The actual type of the target car and microphone for the recognizer will most probably differ from that of the training material which will introduce a certain mismatch. On the other hand for a vehicle from mass production the environment will be constant for a certain car type and defined beforehand. These circumstances suggest to exploit the potential that arises from the knowledge of the environment by adapting a baseline recognizer offline to each target car type.

The HMM adaptation should only consider the environment-related impacts: the noises, the room acoustics of the car and the system response of the microphone. No adaptation to particular speakers should occur so the adaptation material must comprise *many* speakers. Assuming that a speech recognizer in a car must be able to handle several languages and tasks the required adaptation material may sum up to a vast amount. This may be infeasible with real speech recordings in each target environment. So as the basic concept we will focus on the particular approach to use Environment Adapted Databases (EADB) for adaptation. Such EADBs can be easily generated for each car, language and task.

In section 2 the EADB approach is explained and section 3 describes the adaptation methods that will be applied. Investigations under a variety of aspects are then conducted and the experiments and results are finally presented in section 4.

2. ENVIRONMENT ADAPTED DATABASE (EADB)

The EADB approach avoids real speech recordings in the target environment but requires an existing speech database. This database is

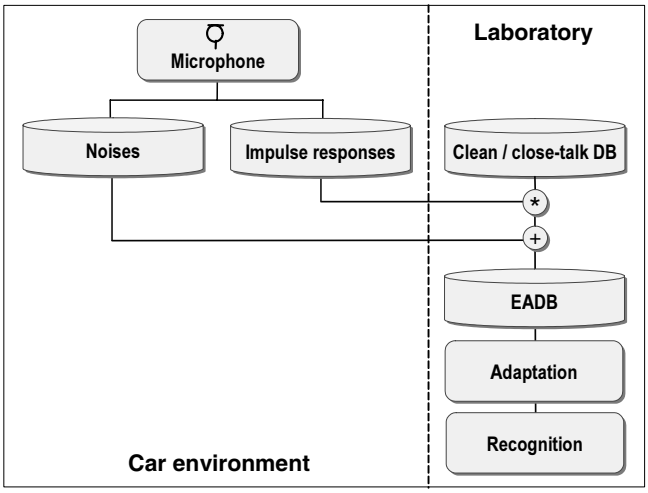

Fig. 1. Scheme for the generation of an Environment Adapted Database (EADB) to be used for HMM adaptation.

then converted into a database characterizing the target car by using impulse response (IR) measurements and driving noise recordings from the target environment according to figure 1. First each sample file of the initial database is convolved with an IR. This incorporates the convolutional influence of the system room-microphone into the signal. Afterwards a section of the noise signals is mixed to the result to superpose the additive disturbances of the environment. The goal is to let the speech files of the produced database "sound" as if they were recorded in the respective car and therefore characterize the specific target environment.

Multi-channel versions of such Environment Adapted Databases (MEADB) were already proposed in [1] for the evaluation of microphone arrays. Similar approaches were also investigated e.g. in [2] and [3] for an office and living room environment respectively. In [4] the method was applied for a car environment but without using measured IRs that turned out in our investigations to be more important for adaptation than the environmental noises.

The input database for the EADB generation is optimally a clean one, i.e. without noise, spectral distortions or reverberation. But such databases are far less usual than the widely available standard car databases for recognition like SpeechDat-Car (SDC; project see [5], databases see [6]). So in the following investigations an approximation is used by resorting to the close-talk channel of such an SDC database. The recordings from SDC are immediately available in huge amounts, various tasks and many languages. An additional

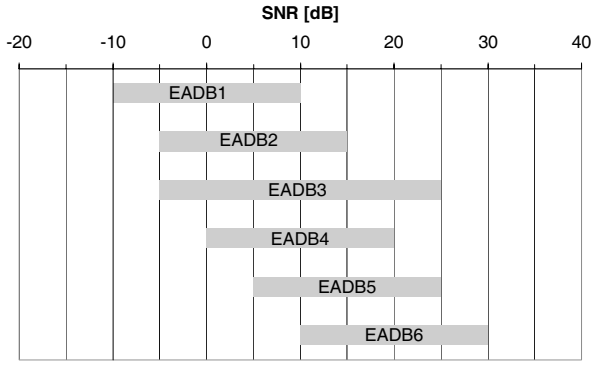

Fig. 2. SNR ranges for the investigated EADBs. Within the limits the step width of possible SNR values is 5 dB.

advantage may be that the recordings already reflect the Lombard effect like it is characteristic for car environments.

2.1. Impulse response measurement

The IRs are measured in the target car using maximum length sequences (MLS, see [7]) that are considered to be the state-of-the-art signals for this purpose. The signals are played back with a loudspeaker and simultaneously recorded from the car environment over the handsfree microphone.

If one measurement is performed it determines the characteristics of the system room-microphone for *one stationary* position of the speaker. In order to reflect the different seat adjustments, body heights and movements of speakers various positions are included in the measurements. The convolution according to figure 1 is then conducted by selecting randomly one IR for each input speech file from the pool of measured IRs. For a large close-talk database the frequency of occurrence for the IRs in the EADB generation process is therefore approximately equally distributed.

2.2. Noise recordings

The noise recordings are done in a car while driving around and should cover all typical noise types like wind, engine, indicator, noise from pedals, noise from passing cars etc. For this purpose a route should be selected that represents a balanced mixture of different driving situations like city traffic, country road and highway.

In the EADB generation process of figure 1 for each input speech file a noise section from the whole noise material is randomly extracted for addition. For a large number of input speech files this means as for the IRs that the use of all the noise material is equally distributed.

2.3. SNR selection

One degree of freedom in the EADB generation process is the chosen SNR because the signal levels of the convolved speech files and the noise files have a random ratio and can arbitrarily be scaled against each other before addition. The approach used for the investigations is to predefine a target minimum and maximum SNR for the resulting EADB and select the SNR for each input file randomly from this range. Figure 2 shows the different SNR ranges that are investigated.

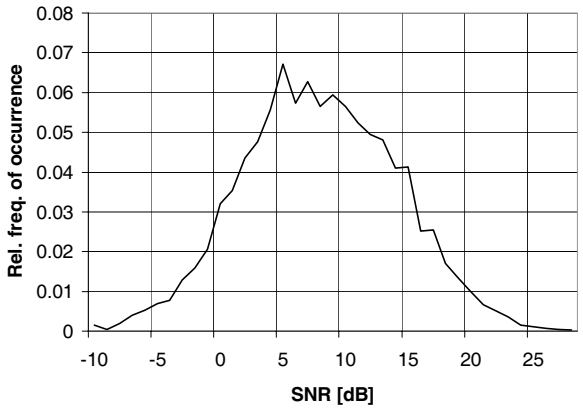

Fig. 3. Histogram showing the SNR distribution of the test database. The mean SNR is 7.8 dB.

3. ADAPTATION METHODS

The adaptation method mainly applied is Maximum Likelihood Linear Regression (MLLR, [8]). Only the mean vectors of the Gaussian densities are transformed with one single affine transformation. Therefore a single regression class covering all densities except those modeling the silence state is used.

In contrast to conventional adaptation scenarios our amount of adaptation material is not very limited due to the use of a whole car database. This permits the investigation of an additional maximum likelihood (ML) training iteration on the baseline HMM as a second approach for adaptation.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

4.1.1. Test material

Especially for the evaluation of the adapted HMMs a test database was collected in the target car and the car's internal microphone was used. The task investigated in this paper are UK English city names[1] and the database comprises 30 speakers (male and female balanced) with a total of 4533 test utterances. The utterances contain 1094 different British city names. The database consists with about equal shares of the following driving situations: car parked – engine running / city traffic / highway. Figure 3 shows the measured SNR distribution for the speech files. In the following sections this test database will be referred to as TestDB.

4.1.2. Baseline recognizer

As baseline a recognizer is trained on the UK English SpeechDat-Car database which consists of speech material from various tasks like phonetically rich sentences, command words, digits, letters, etc. A sampling rate of 8 kHz is used. The baseline HMM serves as the starting point for all adaptation experiments. A total number of about 1200 Gaussian densities model one silence state and strongly tied triphone states ([10]).

[1]The database is available via the Bavarian Archive for Speech Signals ([9]). It contains additionally a spelling task of British city names.

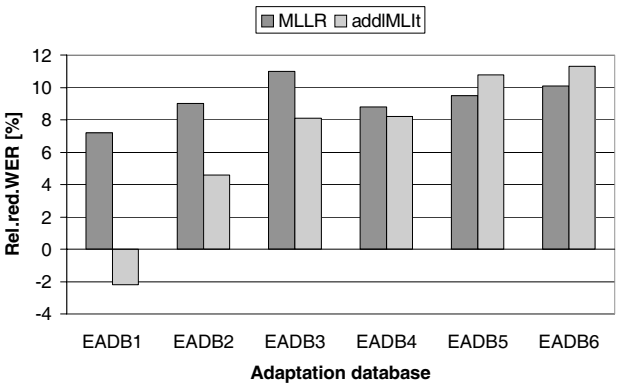

Fig. 4. Performance improvement by car adaptation using the methods MLLR and one additional ML iteration (addlMLIt).

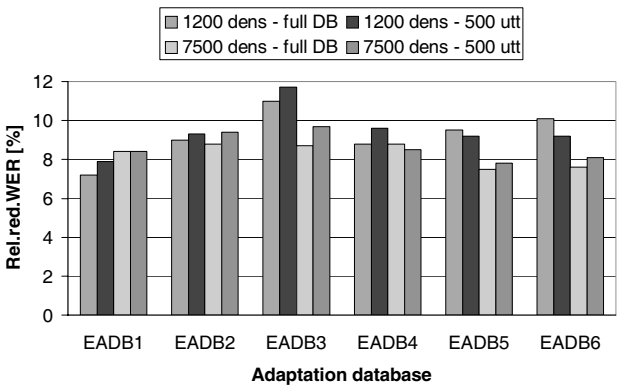

Fig. 5. Performance improvement by MLLR car adaptation with different numbers of adaptation utterances (37857 vs. 500) and densities (1200 vs. 7500).

Only one global variance parameter is used. HMM parameters estimates are based on ML Viterbi training and MCE training with words being the confusion classes ([11]). Each phoneme consists of six states in Bakis-topology. Feature extraction is based on MFCCs and LDA resulting in a feature vector with 24 components. A Wiener Filter is applied for noise reduction and its parameters are optimized during the setup of the baseline recognizer.

For a practical realization of a city names recognizer like in a navigation system it would be a suitable approach to display e.g. a top-five list of the most probable hypotheses. This would considerably increase the absolute recognition accuracies compared to evaluating only the best recognition result. But for the investigations here only the top-one approach will be applied. In this case the word recognition accuracy of the baseline system is 63.4%.

4.2. Variation of EADB properties and adaptation methods

The first experiments investigate the performance improvements for the two adaptation methods MLLR and an additional ML iteration (addlMLIt). Additionally the properties of the EADB used for adaptation are varied according to the parameterizations explained in section 2.3.

Figure 4 shows the results by reporting the relative reductions of word error rate (rel.red.WER). It can be recognized that a significant improvement of up to about 11% (addlMLIt with EADB6, MLLR with EADB3) can be achieved by adaptation and that on average over all EADBs the adaptation methods are comparable. But for each method there are variations of the performance improvement depending on the SNR range of the EADB. The standard deviations of the variations over the EADBs suggest that the method of additional ML iterations (3.2%) is more sensitive to varying properties of the adaptation data than MLLR (1.2%). The other way round it can be concluded that the MLLR adaptation is preferable because it is more robust against a mismatch between an SNR distribution of the EADB (estimated SNR range) and the actual but previously unknown properties of the target environment.

An alternative conclusion could be that both methods are equivalent. But in the case of additional ML iterations the target SNR in the EADB generation process should be set to higher values than expected for the actual recognition environment.

4.3. Variation of amount of adaptation material

In the results from figure 4 all 37857 close-talk utterances that were initially selected from the SDC database for adaptation are included to generate an EADB. But MLLR is known to be able to perform an effective HMM adaptation even with very little data. So in an additional experiment the number of adaptation utterances is reduced to 500 and the results are shown in figure 5 (left two bars of each group of bars) with a variation over the different EADB parameterizations.

The figure illustrates for MLLR adaptation that a limited set of utterances can be employed without loss of performance improvement. This lowers the computation time significantly. Furthermore in case of a language or task where no close-talk data is available yet the effort for collecting clean speech recordings can be reduced. The best result is 11.7% rel.red.WER and is achieved with MLLR, EADB3 and 500 adaptation utterances.

4.4. Variation of HMM size

In the previous experiments the HMM consisted of about 1200 densities. Now it is investigated whether the success of the adaptation depends on the size of the HMM. For this purpose a second baseline HMM is produced with the same properties and training procedure as the first HMM but with about 7500 densities.

The right two bars of each group of bars in figure 5 show the performance improvement for the large HMM using all data and the reduced set of 500 utterances respectively. Note that for each adapted HMM the corresponding baseline HMM with the same number of densities is used to calculate the rel.red.WER. This means that the absolute baseline recognition accuracy is different for the HMMs with 1200 (63.4%) and 7500 densities (72.4%).

It can be recognized that the order of magnitude of the rel.red.WER is comparable for both numbers of densities on average. So the conclusion is that the success of car adaptation using EADBs is not restricted to HMMs of a certain size. This is also true for the reduced set of 500 adaptation utterances.

4.5. Potential of car adaptation

The previous results show a certain capability for performance improvement of a car recognizer by HMM adaptation with EADBs. But a still open issue is to find out how effective the EADB approach

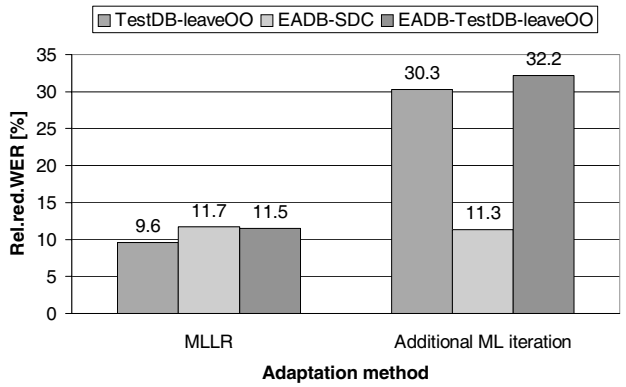

Fig. 6. Performance improvement by adaptation with real speech recordings ("TestDB-leaveOO") compared to SDC-based ("EADB-SDC") and TestDB-based EADBs ("EADB-TestDB-leaveOO"). In case of the EADBs the best result for each method is presented. For the TestDB-based experiments the leave-one-out method is applied ("leaveOO").

works compared to what could be achieved with real speech recordings from the environment.

To determine the potential of adaptation for the given environment an adaptation with the real data of the TestDB can be conducted using the *leave-one-out* method to avoid an overlap of training and test material. But a remaining imponderability is what influence is to be attributed to the different types of utterances in the real database (city names) and the EADBs (mixture of different utterances). The solution is to use also the TestDB as input for the EADB generation and to apply the leave-one-out method again. This can be done because apart from the recordings over the handsfree channel all utterances of the TestDB were recorded simultaneously over the close-talk microphone of a headset.

The results of the different setups are compared for MLLR with the left triple of bars in figure 6. It can be recognized that the performance improvement by HMM adaptation with the EADB approach is comparable and even slightly higher than the improvement by adaptation with real data. And the choice of the initial database for the EADB generation — SDC or the task matching TestDB — does not have a significant impact.

If an additional ML iteration is used as adaptation method the conclusion is different and somewhat surprising as illustrated with the right triple of bars in figure 6. For both leave-one-out experiments, i.e. the adaptation with the TestDB or with the TestDB-based EADB, the performance improvement reaches more than 30% rel.red.WER. This substantially exceeds the results of the adaptation with the SDC-based EADB and of all other previous experiments.

The interpretation is that in this case — apart from the adaptation to the environment — a strong adaptation towards the *task* takes place, i.e. the HMM adapts to the acoustic properties of the city names. It must be remarked that this may be partially attributed to the fact that in our setup all city names of the test set also appear in the adaptation files (of course uttered by different speakers). But anyway the good news is that the same improvement as for the real speech recordings can again be achieved with the EADB approach. It can be concluded that the EADB approach fully exploits the potential for performance improvement in the given environment and therefore proved its suitability for adaptation of car HMMs.

5. SUMMARY AND CONCLUSION

The use of Environment Adapted Databases (EADB) could be proven to be efficient for HMM adaptation to a specific environment as investigated on the example of a car. In our experiments with a city names task the whole potential of HMM adaptation could be exploited for the given environment achieving 11.7% rel.red.WER with MLLR in the generic case and 32.2% performance improvement with task matching material applying an additional maximum likelihood training iteration.

Using EADBs is an approach that reduces the effort for adaptation considerably by avoiding real speech recordings. But for a real product scenario it may still happen that the number of different types and variants of target environments is too large for efficient handling. To reduce the number of variations clustering techniques would be required. Further investigations will be necessary to get answers to this challenge.

6. REFERENCES

[1] H. Höge; J.G. Bauer; C. Geißler; P. Setiawan; K. Steinert, "Evaluation of microphone array front-ends for ASR - an extension of the AURORA framework," in *Int. Conf. on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, 2004, pp. 1611–1614.

[2] O. Gedge; O. Golani; C. Couvreur; K. Linhard, "Evaluation of database adaptation methods," Tech. Rep., SPEECON Deliverable D45, 2002, website of the SPEECON Project: http://www.speechdat.org/speecon/index.html.

[3] V. Stahl; A. Fischer; R. Bippus, "Acoustic synthesis of training data for speech recognition in living room environments," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, 2001.

[4] R. Bippus; A. Fischer; V. Stahl, "Domain adaptation for robust automatic speech recognition in car environments," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, 1999, vol. 5, pp. 1943–1946.

[5] Website of the SpeechDat-Car Project, *http://www.speechdat.org/SP-CAR*.

[6] Website of ELRA, *http://www.icp.grenet.fr/ELRA*.

[7] D.D. Rife; J. Vanderkooy, "Transfer-function measurement with maximum-length sequences," *J. Audio Eng. Soc.*, vol. 37, no. 6, pp. 419–444, June 1989.

[8] C. Leggetter; P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[9] Website of the Bavarian Archive for Speech Signals (BAS), *http://www.phonetik.uni-muenchen.de/Bas/BasKorporaeng.html*.

[10] U. Ziegenhain; J.G. Bauer, "Triphone tying techniques combining a-priori rules and data driven methods," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 2001, pp. 1413–1416.

[11] J.G. Bauer, "On the choice of classes in MCE based discriminative HMM-training for speech recognizers used in the telephone environment," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, 2001, vol. 3, pp. 1633–1636.