ENVIRONMENTAL ADAPTATION WITH A SMALL DATA SET OF THE TARGET DOMAIN

Andreas Maier and Tino Haderlein and Elmar Nöth

University of Erlangen Nuremberg, Chair for Pattern Recognition,
Martenstr.3, 91058 Erlangen, Germany
Andreas.Maier@informatik.uni-erlangen.de

Abstract. In this work we present an approach to adapt speaker-independent recognizers to a new acoustical environment. The recognizers were trained with data which were recorded using a close-talking microphone. These recognizers are to be evaluated with distant-talking microphone data. The adaptation set was recorded with the same type of microphone. In order to keep the speaker-independency this set includes 33 speakers. The adaptation itself is done using maximum a posteriori (MAP) and maximum likelihood linear regression adaptation (MLLR) in combination with the Baum-Welch algorithm. Furthermore the close-talking training data were artificially reverberated to reduce the mismatch between training and test data. In this manner the performance could be increased from 9.9 % WA to 40.0 % WA in speaker-open conditions. If further speaker-dependent adaptation is applied this rate is increased up to 54.9 % WA.

1 Introduction

Gathering training data is a time-consuming and expensive task. Therefore, most speech recognizers are highly specialized to a single recognition task. In most cases the data were collected using a close-talking microphone. However, in real task environments this constraint is often not met. For example car navigation systems often rely on hands-free speaking systems. Close-talking recognizers are not suitable for this task.

We present two techniques in order to adjust a recognizer to a new environment. The first one is MAP and MLLR adaptation [1] in combination with the Baum-Welch algorithm. Therefore, new in-task-domain adaptation data has to be collected. However, the amount of data which is needed for adaptation is much smaller than for training. Since transliterating the new data is even more expensive supervised and unsupervised adaptation are presented in comparison. This procedure is similar to [2]. The second technique is based on the idea to reduce the acoustical mismatch between close-talking training data and the distant-talking evaluation data by producing a "general" reverberation which is convolved into the signal. For training and evaluation we employed the AIBO

published in: Text, Speech and Dialogue (TSD 2006), Proceedings, LNAI 4188, Springer, Berlin, Heidelberg, 2006, pp. 431-437

[3], the Verbmobil [4], and the Fatigue [5] databases which are presented in the following.

2 Databases

2.1 AIBO database

The AIBO Database contains emotional speech of children. In a Wizard-of-Oz experiment children in the age of 12 to 14 years were faced the task to control a Sony AIBO$^{\text{TM}}$ robot [6] by voice. In total 51 pupils (21 male and 30 female) of two different schools were recorded in German language. The whole scenery was recorded by a video camera in order to document the experiment and a head-mounted microphone (*ct*). From the sound track of the video tape a second version of the AIBO corpus was extracted: a distant-talking version (*rm*) was obtained. In this manner no second manual transliteration was necessary because the transcription of the distant-talking and the close-talking version is the same. The distance between the speaker's position and the video camera was approximately 2.5 m. In total 8.5 hours of spontaneous speech data were recorded. The resulting utterances contain 3.5 words each on average. This corpus with 12,858 utterances in total was split into a training, a validation, and a test set with 8,374, 1,310, and 3,174 utterances, respectively. The size of the vocabulary is 850 words plus 350 word fragments. The category-based 4-gram language model which was used for all evaluations was trained on the transcription of the training set and has a perplexity of 50 on the test set.

2.2 Verbmobil (VM)

The Verbmobil (VM) database is a widely used speech collection. We use a German subset of the whole corpus which was already investigated in [7]. The scenario of the corpus is human-human communication with the topic of business appointment scheduling. It contains in total 27.7 hours of continuous speech by 578 speakers of which 304 were male and 274 were female. The size of the vocabulary is 6825 words. On average each of the 12,030 utterances contains 22 words. The size of the training and the test set is 11,714 and 268 utterances, respectively. In order to keep the consistency with earlier experiments the size of the validation set was kept at 48 utterances (cf. [4]). The language model which is employed for the recognition is a category-based 4-gram model which was trained on the transcription of the training data. Its perplexity on the Verbmobil test set is 152.

2.3 Fatigue (FAT)

In order to do an evaluation of the Verbmobil recognizers in acoustical mismatch the Fatigue database which is presented in [5] was used. In this experiment the impact of fatigue on the concentration of participants was investigated. Therefore

3 male and 3 female persons were kept awake a whole night. Among other tasks like playing computer games these six speakers had to read texts which were partially taken from the transcription of the Verbmobil database. The vocabulary and the language model stays the same as in the case of the Verbmobil database. However, this procedure reduces the perplexity of the language model since the test utterances were in the training set of the language model. The perplexity of the 4-gram language model on the Fatigue test set is only 88. The position of the speaker was in front of a microphone array with 15 microphones where 13 were arranged in a linear order. For this work only the central microphone number 7 is used (FAT-Mic7). The distance between the speaker's mouth and the microphone array was approximately 70 cm. The size of the room was 4.5 m × 4.3 m × 3.2 m. This database serves as a reverberated version of parts of the Verbmobil database. So no changes to the recognizer were required.

3 Applied Methods

3.1 Artificial Reverberation

Artificial reverberation is used to create disturbances which resemble those caused by reverberation in a real acoustic environment. It is applied to the signal directly before the feature extraction. So the robustness of the features to reverberation can be improved. The idea is to convolve the speech signal with impulse responses characteristic for reverberation in typical application scenarios e.g. a living room. Thus a reverberated signal can be computed. These impulse responses can be measured in the proposed target environment or generated artificially. For the case of this paper impulse responses have been measured in a specific environment. In current research the artificial reverberation was found to improve the robustness of speech recognizers to acoustic mismatches [5] [8]. In both papers the training data is reverberated using the same twelve impulse responses from assumed speaker positions shown in Fig. 1. The responses differ in the distance, the angle, and the reverberation time T_{60}. The reverberation time is defined as the time that passes until the signal decays to 10^{-6} of its initial sound energy. This corresponds to a reduction of 60 dB. Each response is applied to $\frac{1}{12}$ of the training data. In this manner training data is created which covers a broad variety of possible reverberation.

3.2 Recognizer specifications

The acoustic models of this work are state-tied polyphone models, which are also called semi-continuous Hidden Markov Models (SCHMM). 500 mixture components are shared between all states in the codebook. In addition each state has a weight for each density. In the training the codebook is initialized by the identification of Gaussian mixtures. The transition probabilities between the states are determined by ten iterations of the Baum-Welch algorithm. Then the codebook is re-estimated. The Baum-Welch algorithm and the codebook re-estimation is repeated in an alternating manner for ten times.

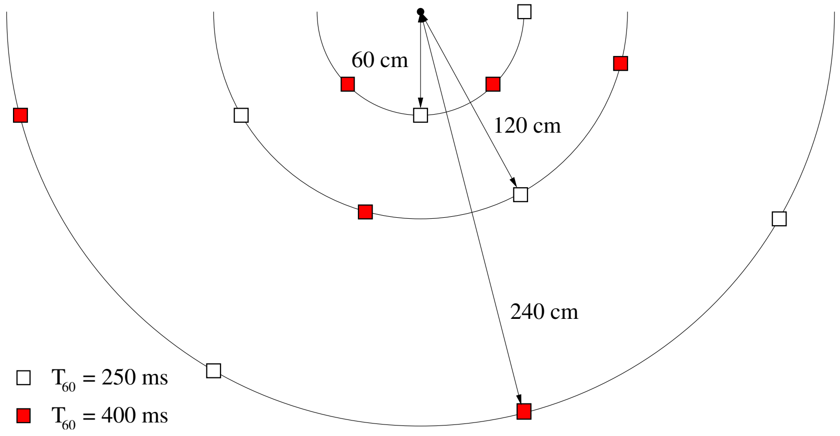

Fig. 1. Recording positions for the impulse responses used in this work (black dot: microphone; squares: assumed speaker positions)

As features the commonly used Mel Frequency Cepstrum Coefficients (MFCC) are employed. We use 12 static features: the spectral energy and 11 cepstral features. Furthermore 12 dynamic features are calculated as an approximation of the first derivative of the static features using a regression line over 5 time frames. The time frames are computed for a period of 16 ms with a shift of 10 ms.

3.3 Adaptation Method

For the adaptation to the new acoustical environment we propose a method using standard codebook adaptation methods – MAP and MLLR – in combination with the Baum-Welch algorithm in order to interpolate the transition probabilities and weights of the acoustic models. As in the case of the training an iterative procedure is applied. First the codebook is adjusted using MAP adaptation. This is followed by MLLR adaptation. Then the transition probabilities and state weights are interpolated using the Baum-Welch algorithm. The experiments showed that three iterations are sufficient in most cases. For this kind of adaptation a transliteration of the adaptation data is necessary and hence the method is supervised. In order to do unsupervised adaptation a transliteration of the adaptation data was generated using the recognizer before each adaptation iteration.

	word accuracy		
		MAP+MLLR adaptation	
training set	baseline	supervised	unsupervised
close-talking (ct)	9.9 %	29.6 %	22.1 %
reverberated (ct rv)	19.4 %	32.9 %	28.4 %
ct + ct rv	16.0 %	33.5 %	27.7 %
$\frac{1}{2}$ ct + $\frac{1}{2}$ ct rv (ct-2)	18.4 %	34.3 %	29.7 %

Table 1. Recognition rates with different sets of artificially reverberated training data with the AIBO recognizer using half of the room microphone data for adaptation

4 Experiments and Results

4.1 AIBO database

Since the distant-talking data of the AIBO database is very noisy and the distance is quite far the recognition on this data set is poor. The plain close-talking recognizer achieves only 9.9 % word accuracy (WA) on the distant-talking evaluation set although the accuracy on the *ct* test set was 79.0 % WA. Training with the distant-talking data achieves 51.1 % WA. This can be seen as an upper boundary for this recognition task. First the effect of artificial reverberation on the adaptation was investigated. Table 1 shows the results which were obtained using only codebook adaptation with MAP followed by MLLR. In this experiment half of the room microphone training data was used for adaptation. *ct* denotes the plain AIBO close-talking recognizer while *ct rv* denotes the recognizer which was trained using only artificially reverberated data. The combination of both was beneficial especially when supervised adaptation was performed. The best results were achieved when half of the training data were reverberated while the other half stayed as is (*ct-2*). Unsupervised adaptation could achieve reasonable recognition rates.

For the case of the *ct-2* training set more investigations concerning the size of the adaptation data and the number of iterations during the adaptation process were done. Fig. 2 shows the results obtained with different sizes of the adaptation data and an increasing number of iterations. As can be seen three iterations of this process seem to be enough. Furthermore $\frac{1}{10}$ of the training data – approximately 30 minutes of speech data – are enough to estimate the new parameters robustly. So the recognition rate can be increased to 40.0 % WA. The best result was obtained with half of the room microphone training data (41.0 % WA). If this method is applied in an unsupervised manner 36.5 % WA can be achieved.

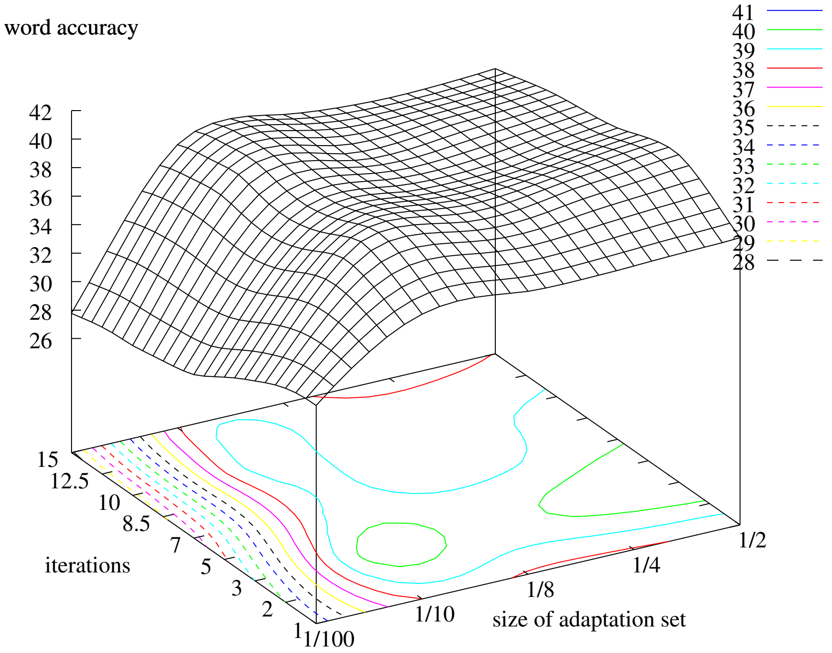

Fig. 2. Supervised Baum-Welch adaptation in combination with MAP and MLLR adaptation of the *AIBO ct-2* recognizer with the *AIBO rm* training set. The word accuracies were obtained with a 4-gram language model.

4.2 Verbmobil and Fatigue database

A similar experiment was done with the Verbmobil database for training and the Fatigue database as test set. The recognizer's accuracy on the close-talking version of the Fatigue test set is 86.9 % WA. Then training was done using the close-talking microphone data (*ct*) and artificially reverberated data (*ct rv*). These recognizers were evaluated using the Fatigue database distant-talking center microphone. Since only six speakers are available in the Fatigue database the adaptation was done in a leave-one-speaker-out (LOO) manner. So adaptation was done using five speakers while evaluation was done with the respectively missing sixth speaker. Thus adaptation data and test data were disjoint. The results displayed in Table 2 show the effect of artificial reverberation in combination with MAP and MLLR adaptation. As can be seen that the adaptation improves the baseline recognizer. However, neither the *ct rv* nor the *ct-2* recognizers can be improved by this speaker-independent adaptation technique.

training set	word accuracy	
	baseline	MAP + MLLR supervised
close-talking (ct)	47.4	57.9
reverberated (ct rv)	71.4	68.8
$\frac{1}{2}$ ct + $\frac{1}{2}$ ct rv (ct-2)	69.2	66.7

Table 2. Recognition rates with different sets of artificially reverberated training data with the Verbmobil recognizer on the FAT-Mic7 test set

5 Discussion

Using MAP and MLLR adaptation and artificial reverberation in a difficult task can improve the recognition a lot. However, many speakers are required. In case of the AIBO database the recognition could be improved from 18.4 % to 40.0 % with the proposed adaptation method using just about 30 minutes of adaptation data (*ct-2* recognizer). However speaker-independent adaptation is not always sensible. The adaptation of the recognizers which were trained using artificially reverberated training data could not achieve further improvement. This is due to the lack of speaker-independency. The five adaptation speakers could not provide speaker-independent adaptation data superior to the artificial reverberation. In addition the respective sixth left out speaker was always of the opposite gender than the majority of the adaptation speakers (3 male and 3 female speakers). So environmental adaptation could only be provided with the plain close-talking (*ct*) recognizer. The use of artificial reverberation, however, was always beneficial.

Further experiments showed that additional speaker-dependent adaptation yields even more improvement on both recognition tasks. In this manner the upper boundary of 51.1 % WA on the AIBO rm task can be broken. So 54.9 % WA on the AIBO database are achieved with the previously adapted recognizer. The speaker-dependent adaptation of the AIBO rm recognizer – and hence the the new upper boundary – is only slightly better with 57.9 % WA. For the Verbmobil database no upper boundary can be determined. Again speaker-dependent adaptation improves the recognition further to 76.1 % WA.

6 Summary

In this paper we showed that speaker-independent adaptation to a certain acoustical environment is possible using a small set of in-task-domain training data. We used three databases in our experiments: the AIBO, the Verbmobil, and the Fatigue database.

Furthermore we added artificial reverberation to the close-talking signal in order to reduce the acoustical mismatch between training and test data. For the adaptation of the codebook we used MAP and MLLR adaptation. The transition probabilities and the state weights were adjusted using the Baum-Welch algorithm.

On the AIBO distant-talking recognition task a total improvement from 9.9 % WA to 40.0 % WA could be done using only about 30 minutes of adaptation data in supervised condition. In unsupervised condition 36.5 % WA could be achieved. However, this kind of speaker-independent adaptation is only sensible if the adaptation set includes enough speakers to keep the constraint of speaker-independency as could be shown with the Fatigue test set. The reverberated version could recognize 71.9 % WA while the speaker-independent adaptation yielded only 68.8 % WA. Speaker-dependent adaptation is suitable for further processing. So 54.9 % WA on the AIBO database and 76.1 % WA on the Fatigue database could be achieved.

References

1. M. Gales, D. Pye, and P. Woodland, “Variance compensation within the MLLR framework for robust speech recognition and speaker adaptation,” in *Proc. ICSLP ’96*, Philadelphia, USA, 1996, vol. 3, pp. 1832–1835.
2. E. Bocchieri, M. Riley, and M. Saraclar, “Methods for task adaptation of acoustic models with limited transcribed in-domain data,” in *Proc. ICSLP ’04*, Jeju Island, Korea, 2004, pp. 326–329.
3. A. Batliner, C. Hacker, S. Steidl, and E. Nöth, “’You stupid tin box’ - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus,” in *Proc. of the 4th International Conference of Language Resources and Evaluation ’04*, Lisbon, Portugal, 2004, pp. 171–174.
4. W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer–Verlag, New York, Berlin, 2000.
5. T. Haderlein, E. Nöth, W. Herbordt, W. Kellermann, and H. Niemann, “Using Artificially Reverberated Training Data in Distant-Talking ASR,” in *Proc. Text, Speech and Dialogue; 8th International Conference, TSD 2005; Karlovy Vary, Czech Republic, 2005*, Berlin, Heidelberg, 2005, vol. 3658 of *Lecture Notes in Artificial Intelligence*, pp. 226–233, Springer–Verlag.
6. Sony, “AIBO Europe – Official Website,” 2005, http://www.aibo-europe.com.
7. G. Stemmer, *Modeling Variability in Speech Recognition*, Ph.D. thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany, 2005.
8. A. Maier, C. Hacker, E. Nöth, and H. Niemann, “Robust parallel speech recognition in multiple energy bands,” in *Pattern Recognition, Proceedings of the 27th DAGM Symposium*, Berlin, Heidelberg, 2005, vol. 3663 of *Lecture Notes in Computer Science*, pp. 133–140, Springer–Verlag.