Using Artificially Reverberated Training Data in Distant-Talking ASR

Tino Haderlein[1], Elmar Nöth[1], Wolfgang Herbordt[2⋆], Walter Kellermann[2], and Heinrich Niemann[1]

[1] University of Erlangen-Nuremberg, Chair for Pattern Recognition (Informatik 5), Martensstraße 3, 91058 Erlangen, Germany
Tino.Haderlein@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de
[2] University of Erlangen-Nuremberg
Chair of Multimedia Communications and Signal Processing
Cauerstraße 7, 91058 Erlangen, Germany

Abstract. Automatic Speech Recognition (ASR) in reverberant rooms can be improved by choosing training data from the same acoustical environment as the test data. In a real-world application this is often not possible. A solution for this problem is to use speech signals from a close-talking microphone and reverberate them artificially with multiple room impulse responses. This paper shows results on recognizers whose training data differ in size and percentage of reverberated signals in order to find the best combination for data sets with different degrees of reverberation. The average error rate on a close-talking and a distant-talking test set could thus be reduced by 29% relative.

1 Introduction

When developing speech-driven human-machine interfaces for hands-free control of devices in a living-room environment, like for television sets and VCRs, the microphones recording the user's utterances will be integrated into the device itself or distributed in the room. This leads to the problem that among other distortions the received signal is reverberated. In our work we used artificially reverberated training data to improve performance of speech recognition in reverberant rooms, as e.g. in [1, 2]. However, we tried to find a training set that is suitable for both reverberated and clean speech and, in general, for unknown target environments. Our research thus aims at ASR systems that are portable between different acoustic conditions. Other well-known methods for improving speech recognition performance on distant-talking data are environment-independent features (see an overview in [3, pp. 39-51]), sometimes with integrated normalization methods as in RASTA-PLP [4], or combining the signals from a microphone array [5]. These were not applied in our experiments.

This paper is organized as follows: In Section 2 we present preliminary experiments with a reduced amount of training data from the EMBASSI corpus [6]

⋆ now with ATR, Kyoto, Japan

published in: Text, Speech and Dialogue (TSD 2005), Proceedings, LNAI 3658, Springer, Berlin, Heidelberg, 2005, pp. 226-233

allowing fast evaluation of different data and recognizer configurations. Based on the findings from these examinations we introduce recognizers with a large training set in Section 3. The data were taken from the VERBMOBIL corpus and allowed us to compare our results with earlier experiments on these data [7]. Section 4 summarizes the results.

2 Preliminary Experiments

2.1 Recognizer Specifications and Baseline System

As in a previous work, we used a baseline recognizer with only one hour of training data for fast evaluation of various setups where the training and test data were taken from the EMBASSI corpus [6]. This German speech collection was recorded in a room with a reverberation time of $T_{60} = 150\,\text{ms}$ (i.e. the time span in which the reverberation decays by 60 dB). It consists of utterances of 20 speakers (10 male, 10 female) who read commands to a TV set and to a VCR, since the topic of EMBASSI was developing speech interfaces for these devices. A close-talking microphone (headset) and a linear array of 11 microphones were used for simultaneous recording. The center of the latter was either 1 meter or 2.5 meters away from the speaker (see Fig. 1). In each one of 10 sessions each speaker read 60 sentences which took approx. between 150 and 180 seconds. The size of the room was $5.8\,\text{m} \times 5.9\,\text{m} \times 3.1\,\text{m}$, the center of the microphone array was at position $(2.0\,\text{m}, 5.2\,\text{m}, 1.4\,\text{m})$. The speaker sat at position $(2.0\,\text{m}, 4.2\,\text{m}, 1.4\,\text{m})$ or $(2.0\,\text{m}, 2.7\,\text{m}, 1.4\,\text{m})$, respectively, i.e. the head was at about the same height as the microphones. The origin of the coordinate system in the room was the left corner behind the speaker.

The training data of the EMBASSI baseline system (EMB-base, see Table 1) consisted of the close-talking recordings of 6 male and 6 female speakers from two sessions (60 min of speech, 8315 words). One male and one female speaker formed the validation set (10 min, 1439 words), and one half of the test set consisted of the remaining three men and three women (30 min, 4184 words). The other half were the corresponding data of the central array microphone, which was 1 m away during one of the used sessions and 2.5 m during the other.

Our speech recognition system is based on semi-continuous HMMs. It models phones in a variable context dependent on their frequency of occurrence and thus forms the so-called polyphones. The HMMs for each polyphone have three to four states. The EMBASSI recognizers have a vocabulary size of 474 words and were trained with a 4-gram language model. For each 16 ms frame (10 ms overlap) 24 features were computed (signal energy, 11 MFCCs and the first derivatives of those 12 static features, approximated over 5 consecutive frames).

Before reverberating the training data artificially we trained a recognizer with EMBASSI data from a distant microphone (EMB-rev) in order to find out which results could maximally be reached when training and test environment were the same. Therefore we used the signals from the microphone from the center of the microphone array whose recordings were synchronously recorded

Table 1. Data sets for the recognizers trained with EMBASSI data ("mic. dist." = microphone distance, "CT art. rev." = close-talking artificially reverberated)

recognizer	training		validation		test	
	mic. dist.	duration	mic. dist.	duration	mic. dist.	dur.
EMB-base	close-talk	60 min	close-talk	10 min	close-talk	30 min
(T_{60}: $\approx$ 0 ms)					1 m	15 min
					2.5 m	15 min
EMB-rev	1 m	30 min	1 m	5 min	*like EMB-base*	
(T_{60}: 150 ms)	2.5 m	30 min	2.5 m	5 min		
EMB-12	close-talk	12·60=	close-talk	12·10=	*like EMB-base*	
(T_{60}: 250, 400 ms)	(artif. rev.)	720 min	(artif. rev.)	120 min		
EMB-2	close-talk+	60 min	close-talk+	10 min	*like EMB-base*	
(T_{60}: 0, 250, 400ms)	CT art. rev.	60 min	CT art. rev.	10 min		

with the close-talking training data. As two EMBASSI sessions were involved, half of the data were recorded at a distance of 1 m and the other half at 2.5 m distance (see Table 1). The situation for the validation data was similar. Only the test data were exactly the same as before.
Table 2 shows that, for distant talkers, the best results are achieved on the reverberated test data, i.e. for those acoustical environments that were present in the training data. For 1 m microphone distance the word accuracy was 94.1% (90.2% on EMB-base) and for 2.5 m distance it was 93.1% (84.1%). The close-talking signals, however, have disadvantages in this approach (87.5% vs. 94.3%). In the table we also added the results for the recognition without a language model in order to show how much the pure acoustic information contributed to the word accuracies. The good results when using the 4-gram model were achieved because the training data were not spontaneous, but read sentences.

Of course training a recognition system with reverberated speech is a simple way to improve the results on test data recorded with a large distance from speaker to microphone. This usually means, however, that the acoustical properties of the training data are the same as in the test data. In a real application the target environment is largely unknown before. Therefore, we investigate in the following to what extent artificially reverberated training data can match various test environments.

2.2 Training the System with Artificially Reverberated Data

If the goal is a recognizer which works robustly in many environments one might suggest that the training data should provide recordings that were made in a lot of different places. This would mean collecting speech data in many rooms with different impulse responses and place the microphone(s) in different angles and distances from the speakers who also have to be available in every location. Reverberating close-talking speech artificially with the help of pre-defined room impulse responses can reduce this effort.

Table 2. Word accuracies for the EMBASSI recognizers (0-gram = no language model)

mic. dist.	lang. model	EMB-base	EMB-rev	EMB-12	EMB-2
close-talk	4-gram	94.3	87.5	91.7	95.5
close-talk	0-gram	70.0	40.0	57.7	71.4
1 m	4-gram	90.2	94.1	94.0	94.4
1 m	0-gram	52.4	66.2	61.9	63.0
2.5 m	4-gram	84.1	93.1	88.4	89.6
2.5 m	0-gram	37.5	63.2	52.4	55.3

The room impulse responses were measured in the room where also the EMBASSI corpus was recorded. However, the reverberation time was changed from $T_{60} = 150\,\text{ms}$ to $T_{60} = 250\,\text{ms}$ and to $T_{60} = 400\,\text{ms}$ by removing sound absorbing carpets and sound absorbing curtains from the room. 12 impulse responses were measured for loudspeaker positions on three semi-circles in front of the microphone array at distances 60 cm, 120 cm, and 240 cm. See Fig. 1 for the experimental setup. The close-talking training data of the baseline recognizer were convolved with each one of the impulse responses separately, i.e. 12 hours of reverberated data (EMB-12; cmp. Table 1) resulted from one hour of close-talking speech.

The results for the recognition experiments are summarized in Table 2. It can be noticed that the recognition performance for the reverberated data increased. Although the acoustical properties of the training data are different from those of the test data, especially for 1 m microphone distance similar results could be achieved as for matching training and testing conditions (94.0% vs. 94.1%). However, the recognition performance for the close-talking test data decreased. So we tested if a mixture of reverberated and clear training data can avoid this problem but still keep the recognition rates for the room microphones on their high level. Therefore we used as one part of the training set the entire training set of the baseline recognizer (see Table 1). The other part consisted of one twelfth of the artificially reverberated training files used in the EMB-12 recognizer, i.e. the new training set (EMB-2; see Table 1) was twice as big as for the baseline system and each room impulse response was present in $\frac{1}{24}$ of the data. Thus the ratio between close-talking and reverberated training data was 1:1. Other ratios are currently being examined.

The results for this approach in Table 2 show that the recognition could be enhanced for all three test sets, even for the close-talking recordings. This is very encouraging in view of a future application, but the question arose if the reason for this improvement was really (only) the reverberation of the training files. Note that the baseline recognizer had a very small training set of about one hour of speech data only, so it might be that the baseline training set (EMB-base) was simply too small for a robust estimation of the phone models. Which percentage of the improvement was the outcome of the sound quality and the size of the data set, resp., had to be estimated during further tests. Furthermore

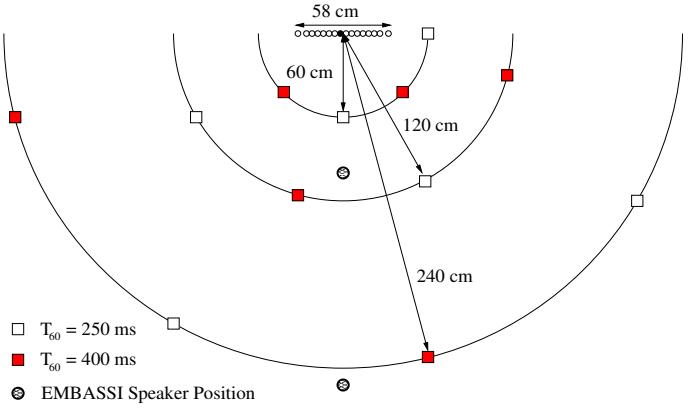

Fig. 1. Assumed speaker positions in the virtual recording rooms for artificially reverberated data; 12 room impulse responses from different positions and with two reverberation times (250 and 400 ms) were used. The circles mark the positions of the speaker in the real EMBASSI recording room

training and test data were both taken from the EMBASSI corpus up to now. The impulse responses for the artificial reverberation of the close-talking signals were measured in the same recording room. In the next section we therefore describe experiments with two other corpora for training and test.

3 Experiments with Verbmobil Training Data

In a next experiment, we study the recognition performance for a larger vocabulary and for longer reverberation time of the testing environment.

3.1 Training and Test Data

A widely used data collection for speech recognition in German-speaking countries is the German part of the VERBMOBIL corpus. We use a subset of this data consisting of about 27.5 hours of close-talking speech produced by 578 speakers (304 male, 274 female; cmp. [7]). The topic in the dialogues is appointment scheduling (spontaneous speech) involving a vocabulary of 6825 words. As test set, we used a subset of a currently unpublished corpus recorded at our faculty which will be denoted as "FAT" in the following. It was recorded in an office room of size $4.5\,\mathrm{m} \times 4.3\,\mathrm{m} \times 3.2\,\mathrm{m}$ with reverberation time $T_{60} = 300\,\mathrm{ms}$. 6 speakers (3 male, 3 female) read transliterations of VERBMOBIL dialogues. Thus the vocabulary of both speech collections was the same and the FAT data could easily serve as test data for the VERBMOBIL recognizers. The distant-talking microphone was placed at the position $(2.0\,\mathrm{m}, 2.5\,\mathrm{m}, 1.4\,\mathrm{m})$, the speaker position was $(2.0\,\mathrm{m}, 1.5\,\mathrm{m}, 1.4\,\mathrm{m})$, i.e. 1 m away from the distant-talking microphone.

The origin of the coordinate system in the room was the left corner behind the speaker.

The training and validation data for the baseline VERBMOBIL recognizer (VM-base) were the same set as in [7]. As in the previous experiments, the recognizer was trained on mixtures of clean and artificially reverberated close-talking signals. The important difference, however, is that the sizes of training and validation set were not changed for the different acoustic conditions. Thus the changes in the results are only dependent of the degree of reverberation in the data, because the acoustic model of a specific phone gets the same amount of training data in all training processes, only the acoustic conditions change.

Concerning the training set, three different recognizers were set up (Table 3) comparable to those with the EMBASSI data:

- **VM-base:** This is the baseline VERBMOBIL recognizer as described in [7]. It was trained with close-talking recordings only (257,810 words, 11714 utterances).
- **VM-12:** All close-talking recordings were reverberated. The impulse responses were changed for each utterance for preventing that all utterances from the same speaker are convolved with the same impulse responses.
- **VM-2**: As for EMB-2 (Table 1) half of the training set consisted of close-talking signals and half of reverberated files. The 12 room impulse responses were equally distributed over the utterances.

The fact that only 48 utterances were in the original VERBMOBIL validation set was inconvenient for the test series as each one of the 12 room impulse responses was represented in the validation lists of VM-12 and VM-2 by very few files. Nevertheless the file lists were not changed in order to get results comparable with experiments in [7].

The recognizers were evaluated on four data sets (cmp. Table 3):

- the original VERBMOBIL test set (268 close-talking recordings, 4781 words, 30 min of speech) as defined in [7].
- the artificially reverberated VERBMOBIL test set: The original test set was convolved with the same 12 room impulse responses also used for the corresponding training data. The 268 files contain the 12 room impulse responses with equal proportions.
- the FAT close-talking set: The 1445 files contain 24738 words (vocabulary size: 865 words) and have a total duration of 150 min.
- the FAT room microphone set: These data were synchronously recorded with the close-talking data by a room microphone 1 m away from the speaker. This microphone was of the same type as those used for the EMBASSI corpus.

As the texts read in the FAT test data were transliterations of VERBMOBIL dialogues, all the utterances were in the training data of the language model. Therefore, the recognition results which are obtained for the FAT close-talking data using the 4-gram language model are better than for the (non-overlapping) VERBMOBIL close-talking test set (see Table 4). The test set perplexity of the FAT data was 87.7 while for the VERBMOBIL language model test data it was 151.5.

Table 3. Data sets for the recognizers trained with VERBMOBIL data ("mic. dist." = microphone distance, "CT art. rev." = close-talking artificially reverberated)

recognizer	training mic. dist.	training dur.	validation mic. dist.	validation dur.	test mic. dist.	test dur.
VM-base (T_{60}: ≈ 0 ms)	close-talk	27 h	close-talk	7 min	close-talk	30 min
					CT art. rev.	30 min
					FAT CT	150 min
					FAT 1 m	150 min
VM-12 (T_{60}: 250, 400 ms)	close-talk (artif. rev.)	27 h	close-talk (artif. rev.)	7 min	*like VM-base*	
VM-2 (T_{60}: 0, 250, 400 ms)	close-talk+ CT art. rev.	13.5 h 13.5 h	close-talk+ CT art. rev.	3.5 min 3.5 min	*like VM-base*	

Table 4. Word accuracies for the VERBMOBIL recognizers (0-gram = no language model)

test set	lang. model	VM-base	VM-12	VM-2
VERBMOBIL close-talk	4-gram	80.1	72.1	77.9
VERBMOBIL close-talk	0-gram	51.4	37.4	49.1
VERBMOBIL art. rev.	4-gram	59.9	67.5	67.4
VERBMOBIL art. rev.	0-gram	28.5	39.8	37.6
FAT close-talk	4-gram	86.8	81.6	85.5
FAT close-talk	0-gram	49.4	38.3	46.5
FAT reverb.	4-gram	47.8	71.3	69.4
FAT reverb.	0-gram	12.5	32.3	28.8

3.2 Results

Table 4 summarizes the results on the VERBMOBIL based recognizers. The word accuracy for the FAT close-talking test set is the highest for the VM-base recognizer (86.8% word accuracy using a 4-gram language model) and lowest for VM-12 where only reverberated data was in the training set (81.6%). VM-2 almost reaches the baseline result (85.5%). Regarding the FAT data recorded at 1 m distance in a room with $T_{60} = 300$ ms the close-talking recognizer VM-base shows least accuracy as expected (47.8%) and VM-12 the highest one (71.3%). Here VM-2 with 69.4% also nearly reaches the same value. Taking the average of the results on FAT close-talking data and distant-talking data the baseline word accuracy of 68.3% can be improved by 29.0% relative to a word accuracy of 77.5% on VM-2 (VM-12 reaches 76.5%). This result shows that artificially reverberated training data can help to improve the robustness of speech recognition in reverberant acoustic environments for mismatch of the room impulse responses for training and testing.

4 Conclusions and Outlook

We tested artificially reverberated training data for improving the robustness of ASR against reverberation. For training, we used a small subset of the EMBASSI and the German VERBMOBIL corpus, respectively, using room impulse responses from environments with $T_{60} = 250\,\text{ms}$ and 400 ms reverberation time. For testing, we used the test set of the FAT corpus which contains synchronously recorded signals from a close-talking microphone and a distant-talking microphone at a distance of 1 m in a room with $T_{60} = 300\,\text{ms}$ reverberation time. The average word accuracy for both test subsets on a VERBMOBIL recognizer trained with close-talking data (VM-base) was 68.3%. Training with artificially reverberated data (VM-12) lead to an increase for reverberated data but to a decrease for close-talking data. Using half of both training sets in another recognizer (VM-2) did not only give the best average result (77.5%), but with merely moderate loss on the single subsets. Future experiments will include optimizing the relation between close-talking and distant-talking training data and testing other kinds of features like our MFCC variant with μ-law companded Mel spectrum coefficients [6].

Acknowledgments

Our work was partially supported by the German Federal Ministry of Education and Research (grant no. 01 IMD 01 F) in the frame of the SmartWeb project. The responsibility for the contents of this study lies with the authors.

References

1. L. Couvreur, C. Couvreur, and C. Ris. A Corpus-Based Approach for Robust ASR in Reverberant Environments. In *Proc. of International Conference on Spoken Language Processing (ICSLP)*, volume 1, pages 397–400, Beijing, China, 2000.
2. V. Stahl, A. Fischer, and R. Bippus. Acoustic Synthesis of Training Data for Speech Recognition in Living Room Environments. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 21–24, Salt Lake City, Utah, 2001.
3. J.-C. Junqua. *Robust Speech Recognition in Embedded Systems and PC Applications*. Kluwer Academic Publishers, Boston, 2001.
4. B.E.D. Kingsbury and N. Morgan. Recognizing Reverberant Speech with RASTA-PLP. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1259–1262, Munich, Germany, 1997.
5. M. Omologo, P. Svaizer, and M. Matassoni. Environmental conditions and acoustic transduction in hands-free speech recognition. *Speech Communication*, 25(1–3):75–95, 1998.
6. T. Haderlein, G. Stemmer, and E. Nöth. Speech Recognition with μ-Law Companded Features on Reverberated Signals. In V. Matoušek and P. Mautner, editors, *Proc. 6th Int. Conf. on Text, Speech and Dialogue – TSD 2003*, volume 2807 of *Lecture Notes in Artificial Intelligence*, pages 173–180, Berlin, 2003. Springer–Verlag.
7. G. Stemmer. *Modeling Variability in Speech Recognition*. PhD thesis, Chair for Pattern Recognition, University of Erlangen-Nuremberg, Germany, 2005.