Experiments in Distant Talking Speech Recognition Using a Standard Database

Gernot A. Fink, Sascha Hohenner

Universität Bielefeld, Technische Fakultät, 33564 Bielefeld, Deutschland, Email: {gernot,sascha}@techfak.uni-bielefeld.de

Introduction

Distant talking speech recognition is applied in situations where the use of close-talking microphones is not feasible, e.g. when communicating with a mobile robot. The data is captured with multiple microphones from some distance to the talker. However, not only the desired speech but also sound from interfering sources is picked up. Additionally, the received signals are corrupted by echoes introduced by the acoustic environment.

Therefore, a degradation in recognition quality can be observed for distant talking applications compared to results obtained on "clean" speech. Various methods, as e.g. beam-forming and adaptation techniques, have been proposed for obtaining the maximum possible recognition quality even in such adverse environments. However, a direct comparison of results is usually not possible as larger speech databases recorded in both a distant talking and a clean-speech configuration are not available. In this paper we propose a novel solution to this problem.

Related Work

A rather straight-forward solution to the problem is to take into account the distant-talking setup already when collecting a speech database. This approach was for example pursued in [4]. In an office environment with some noise sources present speech prompts uttered by several subjects were recorded both with a close-talking microphone and a microphone array. As data collection is always a considerable effort especially with more elaborate setups, the richness of the data collected is rather limited. Only connected digits were uttered and only by 24 speakers. Furthermore, the baseline performance on the clean data can not be compared directly to results reported in the literature.

Therefore, in [6] the clean data was taken from a standard speech database – namely the *Wall-Street Journal* corpus (WSJ0) [5]. In a setup where 4 microphones were mounted on a PDA subjects uttered text prompts from the WSJ data which were recorded via the PDA array. For this configuration the baseline performance can easily be compared to results from the literature. However, the test data recorded via the PDA are still only loosely related to the standard benchmarks. Additionally, as only the test data is available as distant-talking speech the possible improvement in performance for matched training and testing conditions can not be assessed.

Corpus Design

In order to be able to clearly assess the performance of our approach for distant talking speech recognition developed for our mobile robot BIRON with respect to the ideal close-talking scenario, we decided to make a standard speech database (the WSJ0 corpus [5]) usable for a comparative study. We simulated a talker using a high-quality studio loudspeaker (GENELEC 1019A with neutral frequency characteristics) set up at an appropriate distance to our robot for "uttering" - i.e. re-playing - the complete corpus. This data was then recorded via the stereo-microphones of the robot yielding a distant talking version of the original corpus.

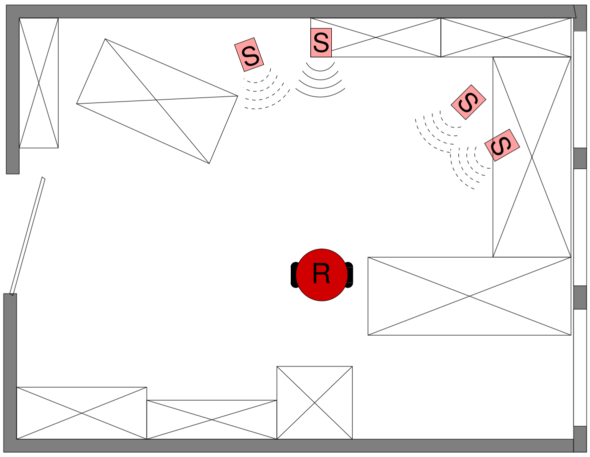

Figure 1: Setup for re-recording the clean speech data in an office environment at different source positions S with stereo-microphones mounted on the robot R.

The setup for creating the distant-talking version of the WSJ0 corpus is shown in Fig. 1. Our mobile robot BIRON was positioned approximately in the middle of an office room with moderate reverberation ($T_{60} \simeq 270$ ms). The robot's two microphones have a distance of 28 cm and reach a height of 106 cm. The simulated speech sources S were positioned at a height of 157 cm with 4 different relative orientations (0°, -20°, +45°, +60°) at a distance of 150 cm to the robot R.

With this configuration the speaker independent training data (74 speakers, approx. 15 hours of speech) and the data for the 5k closed vocabulary test (approx. 40 minutes) was re-recorded with the stereo microphone setup without additional noise (WSJ0-Stereo) and with interfering noise from the robot itself (WSJ0-BIRON) caused by the two on-board computers and the activated laser scanner. The training data was re-recorded in the frontal

*The work described in this paper was supported by the German Research Foundation (DFG) within the Collaborative Research Centre "Situated Artificial Communicators" (SFB 360).

configuration only (0°) where no significant time-delay could be observed in the different recording channels. The test set, however, was re-recorded with the simulated speaker placed at all four different relative positions to the robot.

Recognition Systems

The speech recognition systems used for the experiments were developed within the ESMERALDA framework [1].

The features computed are a variant of the well known mel-frequency cepstral coefficients. We use a sampling rate of 16 kHz, 16 ms frames and a 10 ms frame rate. Variations in the recording channel and to some extent also the acoustic environment are compensated by a causal cepstral mean normalization. In order to increase the robustness of the speech recognizer to noise we integrated a model of forward masking into the feature extraction process [7]. For acoustic modelling we use semi-continuous Hidden Markov Models with tri-phone sub-word units and linear topology. The transcriptions for the WSJ data were derived from the "Carnegie Mellon Pronouncing Dictionary" (v0.6). In order to optimally exploit the training data we apply a data-driven state clustering which creates approximately 5800 model states from the speaker independent training data. For the experiments reported below we used a 5k lexicon (closed vocabulary test) and a bi-gram language model.

For applying the recognition systems to the stereo speech-data recorded with the robot's microphones we use delay-and-sum beamforming (cf. [3]). The merging of the individual channels is performed in the spectral domain in order to be able to apply channel normalization and forward masking separately for each channel [2].

Results

In the experimental evaluations we investigated three major aspects of distant-talking speech recognition: The degree of performance degradation when moving from close-talking to distant-talking data, the impact of matched training conditions, and the influence of the relative position between speech source and microphones.

Training: WSJ0-Mono				
Test	Mono	Stereo	Left	Right
WSJ0-Mono	12,1%	-	-	-
WSJ0-Stereo		33,8%	38,3%	36,2%
WSJ0-BIRON		65,7%	67,1%	79,4%

Table 1: Word error rates (WER) achieved when training on close-talking data and testing on distant-talking speech without and with additional noises of the robot.

Table 1 shows the recognition results obtained when using a recognizer trained on the original close-talking WSJ0 data (WSJ0-Mono) and testing on distant-talking speech. Though beamforming improves the performance compared to testing on a single distant-talking channel the word error rate (WER) almost triples when moving from close-talking to distant-talking speech.

Training: WSJ0-Stereo			
Test	Stereo	Left	Right
WSJ0-Stereo	22,2%	25,3%	26,6%
WSJ0-BIRON	27,1%	29,8%	29,8%

Table 2: Word error rates (WER) achieved when both training and testing on distant-talking speech, i.e. for matched conditions.

A considerable improvement can be obtained in the matched condition with both training and testing on distant-talking data as shown in table 2. In this configuration even the additional noises of the robot have only little impact on the recognition quality.

Training: WSJ0-Stereo				
Test	0°	-20°	+45°	+60°
WSJ0-Stereo	22,2%	22,7%	24,2%	27,3%
WSJ0-BIRON	27,1%	27,8%	30,1%	32,9%

Table 3: Word error rates (WER) achieved on distant-talking speech for different relative orientations of speech source and robot.

Finally, table 3 shows that there is a slight degradation in performance with an increasing deviation from the frontal recording condition.

In recent informal user experiments with our mobile robot BIRON it could also be observed that the distant-talking recognizer achieves good performance and robustness on real-world data.

References

[1] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

[2] S. Hohenner. *Spracherkennung für agierende Systeme.* PhD thesis, Technische Fakultät, Universität Bielefeld, 2005.

[3] S. J. Leese. Microphone arrays. In G. M. Davis, editor, *Noise Reduction in Speech Applications*, chapter 7, pages 179–197. CRC Press, Boca Raton, 2002.

[4] M. Matassoni, M. Omologo, D. Giuliani, and P. Svaizer. Hidden Markov model training with contaminated speech material for distant-talking speech recognition. *Computer Speech & Language*, 16(2):205–223, 2002.

[5] D. B. Paul and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Speech and Natural Language Workshop*. Morgan Kaufmann, 1992.

[6] M. L. Seltzer, B. Raj, and R. M. Stern. Likelihood-maximizing beamforming for robust hands-free speech recognition. *IEEE Trans. on Speech and Audio Processing*, 12(5):489–498, 2004.

[7] S. Wendt, G. A. Fink, and F. Kummert. Forward masking for increased robustness in automatic speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 615–618, Aalborg, 2001.