Automated vocal assistant: distant microphone preprocessing

B. Van Den Broeck[1] , P. Karsmakers[1,2], K. Demuynck[2], H. Van hamme[2], B. Vanrumste[1,2]

[1]MOBILAB: Biosciences and Technology Department, K.H.Kempen, Belgium

[2]ESAT, K.U.Leuven, Belgium

Keywords: Vocal assistance - Distant microphones

1. Introduction

Many simple tasks such as opening/closing the curtains or even changing channels on the TV, can be quite a challenge for elderly or disabled persons and even might engender risks such as falling. Nowadays these persons often need human assistance . An automated vocal assistant could be the solution for helping these people in leading an independent live at a reasonable economical cost.

The recognition rate of automatic speech recognition of vocal commands strongly depends on the acquired signal quality, which decreases with the distance between mouth and microphone. Therefore a close-talk microphone, e.g. mounted on a headset, is the preferred option from a technical point of view. However, users may not accept or forget to wear headsets or perform activities that are incompatible with wearing a headset such as sleeping. In those cases distant microphones combined with speech enhancement preprocessing can reduce the noise level in the speech signal and remove most of the reverberation. In this work a state-of-the-art multiple-channel speech-enhancement, namely spatial preprocessed speech distortion rated multichannel wiener filter or SP-SDR-MWF [1], is combined with a state-of-the-art speech-recognizer built with the SPRAAK toolkit [2]. The performance of this combination is evaluated in terms of and tested for Word Error Rates (WER).

Furthermore, the acceptability of such a system strongly depends on the overall cost. Therefore electret microphones are used, which come at prices of about 1-2 € a piece.

2. Speech enhancement

When using distant microphones, a spatial preprocessing step such as delay-and-sum beamforming, is often used because of its simplicity and ability to boost the signal quality. This normally requires matched microphones and amplifiers, which increases the cost. The SP-SDR-MWF algorithm was selected for the speech enhancement because of its ability to work with unmatched microphones. SP-SDR-MWF introduces a parameter μ for balancing the effects of unmatched microphones to the obtained noise reduction.

The SPRAAK speech recognizer was preconfigured using noise robust acoustic models trained on read speech recorded with a close-talk microphone, and was, in this work, treated as a black-box that converts speech into word strings and returns the corresponding WER.

3. Experiments and results

A microphone-array was built out of five microphones evenly spaced at 3cm. A speaker was placed in front of the array at a distance of 4m. The speech used was a playback of the noise free part of the Aurora4 dataset [3].

Two experiments were performed: a) in a noisy environment, with white noise playing at a distance of 2.2m and an angle of 26° relative to the speech source and b) the same environment with the noise source turned off. Both recordings where processed using various values of μ, and the resulting "clean" speech was processed by the speech recognizer which returned the WER shown in Figure 1. The green lines are the mean WER obtained by the unprocessed microphone outputs and the blue lines are the WER after processing. It is clear that the resulting WER are lower after the processing. More interestingly, it is noticed that the results with various μ show an optimum, which means that there exists an optimum balance between the effects of microphone mismatch and noise reduction. Furthermore, the effect of processing multiple channels is greater in the noisy environment (about 20% gain compared to the unprocessed microphone outputs) than in the less noisy environment (about 2% gain).

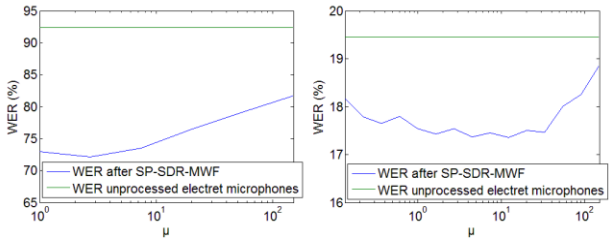

Figure 1: WER in function of μ: left) noisy, right) quantify less noisy

4. Conclusion and future work

The results presented here for a noisy environment probably aren't yet adequate for the use in a vocal assistant. But these results can be improved by ensuring that there are microphones placed closer to the speaker. To do so for random speaker positions, many cheap and small microphone units, operating in a wireless sensor network, could be evenly distributed in a living room.

References

[1] A. Spriet, M. Moonen, J. Wouters. "Stochastic gradient implementation of spatially pre-processed Multi-channel". 2003.

[2] www.spraak.org

[3] N. Parihar, J. Picone, D. Pearce, H.G. Hirsch. "Performance analysis the Aurora large vocabulary baseline system.", Proceedings of the European Signal Processing Conference, Vienna, Austria, 2004