Abstract M. Sc. Jahn Heymann

Robust multichannel ASR with joint optimization of the front- and back-end

Abstract:

The performance of an automatic speech recognition system can be significantly improved by using multiple microphones and exploiting spatial information. Although some approaches feed the raw audio signals directly to the acoustic model, most systems rely on classical signal processing to combine and filter the different channels. This talk focuses on two techniques which have shown to be particularly effective in this scenario: Statistical beamforming and dereverberation with the weighted predictive error (WPE) method. Both are augmented with a neural network to estimate the necessary signal statistics leading to a powerful model while preserving the advantages of a linear filter.

Experiments on the CHiME and REVERB challenge data as well as on a large-scale voice search dataset underline the performance improvements in terms of word error rates for both techniques. Further, the signal processing models are optimised jointly with the acoustic model by backpropagating the gradients through the filter operations, effectively turning the system into one big acoustic model with a special structure tailored towards signal enhancement. The advantages, performance, training strategies and issues, but also potential drawbacks of this joint approach are reviewed and discussed.