Topic 1: Bimodal Speech Recognition

Audio-visual speech recognition makes use of extra video data, in particular lip-reading information, to improve the performance of a traditional acoustic-only speech recognizer. Based on state-of-the-art speech recognition technology, information fusion between acoustic and visual cues is attempted on the level of phones, where the visual stimulus is described by means of an underspecified phone representation. These two representations are then to be combined by a third component for word recognition.

As an alternative to conventional feature processing techniques, Articulatory Features can be used as an intermediate representation, capturing relevant characteristics of the speech production information. In a two-level AVSR system, we applied Hidden Markov Models (HMM) for modelling abstract articulatory classes, and then designed an N-best decision schema to decide the best articulatory feature tuples, in order to achieve a better recognition performance. We also try to make use of other machine learning models (like ANN, or hybrid ANN/HMM) for processing multi-stream information fusion problems.

The questions addressed in this proposal also include how the recognition results are influenced by accidentally missing information from one of channels.