Audio-Visual Dictionary Learning and ProbabilisticTime-Frequency Masking in Convolutive and Noisy Source Separation

Dr. Wenwu Wang, LSSC Consortium

In existing audio-visual blind source separation (AV-BSS) algorithms, the AV coherence is usually established through statistical modelling, using e.g. Gaussian mixture models (GMMs). These methods often operate in a low-dimensional feature space, rendering an effective global representation of the data. The local information, which is important in capturing the temporal structure of the data, however, has not been explicitly exploited. In this talk, we present a new method for capturing such local information, based on audio-visual dictionary learning (AVDL). We address several challenges associated with AVDL, including cross-modality differences in size, dimension and sampling rate, as well as the issues of scalability and computational complexity. Following a commonly employed bootstrap coding-learning process, we have developed a new AVDL algorithm which features, a bimodality balanced and scalable matching criterion, a size and dimension adaptive dictionary, a fast search index for efficient coding, and cross-modality diverse sparsity. We also show how the proposed AVDL can be incorporated into a BSS algorithm. As an example, we consider binaural mixtures, mimicking aspects of human binaural hearing, and derive a new noise-robust AV-BSS algorithm by combining the proposed AVDL algorithm with Mandel's BSS method, which is a state-of-the-art audio-domain method using time-frequency masking. We have systematically evaluated the proposed AVDL and AV-BSS algorithms, and show their advantages over the corresponding baseline methods, using both synthetic data and visual speech data from the multimodal LILiR Twotalk corpus.