Using visual and auditory cues for audio enhancement

Abstract:

In this talk I will present two recent works:

1) Dynamic Temporal Alignment of Speech to Lips

Many speech segments in movies are re-recorded in a studio during post-production, to compensate for poor sound quality as recorded on location. Manual alignment of the newly-recorded speech with the original lip movements is a tedious task. We present an audio-to-video alignment method for automating speech to lips alignment, stretching and compressing the audio signal to match the lip movements. This alignment is based on deep audio-visual features, mapping the lips video and the speech signal to a shared representation. Using this shared representation we compute the lip-sync error between every short speech period and every video frame, followed by the determination of the optimal corresponding frame for each short sound period over the entire video clip. We demonstrate successful alignment both quantitatively, using a human perception-inspired metric, as well as qualitatively. The strongest advantage of our audio-to-video approach is in cases where the original voice is unclear, and where a constant shift of the sound can not give perfect alignment. In these cases, state-of-the-art methods will fail.

2) Neural separation of observed and unobserved distributions

Separating mixed distributions is a long standing challenge for machine learning and signal processing. Most current methods either rely on making strong assumptions on the source distributions or rely on having training samples of each source in the mixture. In this work, we introduce a new method - Neural Egg Separation - to tackle the scenario of extracting a signal from an unobserved distribution additively mixed with a signal from an observed distribution. Our method iteratively learns to separate the known distribution from progressively finer estimates of the unknown distribution. In some settings, Neural Egg Separation is initialization sensitive, we therefore introduce Latent Mixture Masking which ensures a good initialization. Extensive experiments on audio and image separation tasks show that our method outperforms current methods that use the same level of supervision, and often achieves similar performance to full supervision.