Audio-visual speech and singing voice separation

Source separation is the automatic estimation of the individual isolated sources that make up the audio mixture. The goal of this project is to separate a human voice in a mixture by using both the audio and video modalities. We are interested in both speech and singing voice signals. The most direct applications of speech separation are speaker identification and speech recognition (for example, to create automatic captioning of videos). While some of the applications of singing voice separation are: automatic creation of karaoke, music transcription, or music unmixing and remixing. Leveraging visual and motion information from the target person's face is particularly useful when there are different voices present in the mixture. Deep neural networks that extract features from the video sequence will be explored and used in conjunction with an audio network in order to improve the audio source separation task by incorporating visual cues. An unsupervised model will be explored.

Related demos:

https://ipcv.github.io/VoViT/demos/
https://ipcv.github.io/Acappella/demos/

Contact: Gloria Haro, gloria.haro@upf.edu