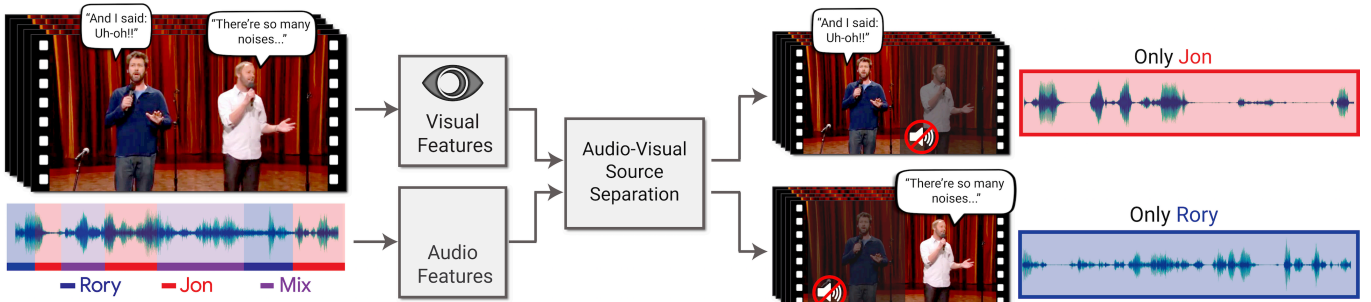

(a) Input video frames and audio (b) Processing (c) Output clean audio for each speaker (our result)