מכון ויצמן למדע

THE WEIZMANN INSTITUTE OF SCIENCE

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Vision and Robotics Seminar

Room 1 ,Ziskind Building
on Thursday, Jun 28, 2018
at 12:15

Ariel Ephrat
The Hebrew University of Jerusalem

Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation

Abstract:

We present a joint audio-visual model for isolating a single speech signal from a mixture of sounds such as other speakers and background noise. Solving this task using only audio as input is extremely challenging and does not provide an association of the separated speech signals with speakers in the video. In this paper, we present a deep network-based model that incorporates both visual and auditory signals to solve this task. The visual features are used to "focus" the audio on desired speakers in a scene and to improve the speech separation quality. To train our joint audio-visual model, we introduce AVSpeech, a new dataset comprised of thousands of hours of video segments from the Web. We demonstrate the applicability of our method to classic speech separation tasks, as well as real-world scenarios involving heated interviews, noisy bars, and screaming children, only requiring the user to specify the face of the person in the video whose speech they want to isolate. Our method shows clear advantage over state-of-the-art audio-only speech separation in cases of mixed speech. In addition, our model, which is speaker-independent (trained once, applicable to any speaker), produces better results than recent audio-visual speech separation methods that are speaker-dependent (require training a separate model for each speaker of interest).

Joint work with Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, Bill Freeman and Miki Rubinstein of Google Research.