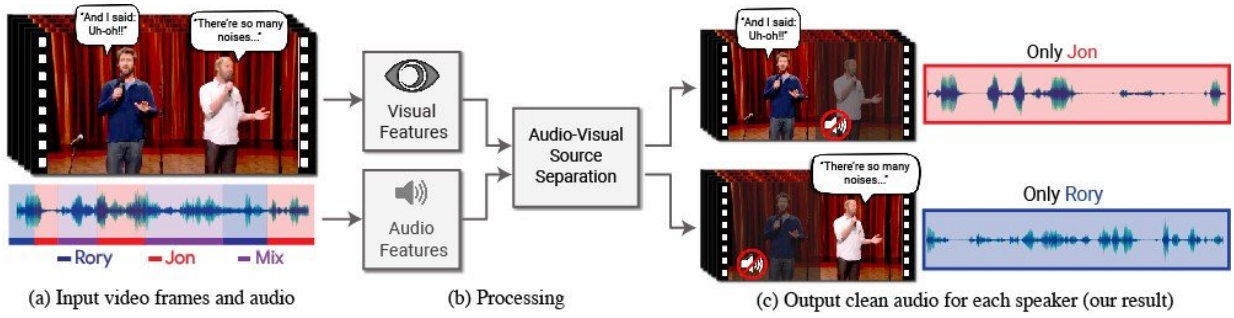
New method enables high quality speech separation

June 5 2018



A new model isolates and enhances the speech of desired speakers in a video. (a) The input is a video (frames + audio track) with one or more people speaking, where the speech of interest is interfered by other speakers and/or background noise. (b) Both audio and visual features are extracted and fed into a joint audio-visual speech separation model. (c) The output is a decomposition of the input audio track into clean speech tracks, one for each person detected in the video. The speech of specific people is enhanced in the videos while all other sound is suppressed. The new model was trained using thousands of hours of video segments from the team's new dataset, AVSpeech, which will be released publicly. Credit: Authors/Google Video stills: Courtesy of Team Coco/CONAN

People have a natural knack for focusing on what a single person is saying, even when there are competing conversations in the background or other distracting sounds. For instance, people can often make out what is being said by someone at a crowded restaurant, during a noisy party, or while viewing televised debates where multiple pundits are

talking over one another. To date, being able to computationally—and accurately—mimic this natural human ability to isolate speech has been a difficult task.

"Computers are becoming better and better at understanding speech, but still have significant difficulty understanding speech when several people are speaking together or when there is a lot of noise," says Ariel Ephrat, a Ph.D. candidate at Hebrew University of Jerusalem-Israel and lead author of the research. (Ephrat developed the new model while interning at Google the summer of 2017.) "We humans know how to understand speech in such conditions naturally, but we want computers to be able to do it as well as us, maybe even better."

To this end, Ephrat and his colleagues at Google have developed a novel audio-visual model for isolating and enhancing the speech of desired speakers in a video. The team's deep network-based model incorporates both visual and auditory signals in order to isolate and enhance any speaker in any video, even in challenging real-world scenarios, such as video conferencing, where multiple participants oftentimes talk at once, and noisy bars, which could contain a variety of background noise, music, and competing conversations.

The team, which includes Google's Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein, will present their work at SIGGRAPH 2018, held 12-16 August in Vancouver, British Columbia. The annual conference and exhibition showcases the world's leading professionals, academics, and creative minds at the forefront of computer graphics and interactive techniques.

In this work, the researchers did not just focus on auditory cues to separate speech but also visual cues in the video—i.e., the subject's lip movements and potentially other facial movements that may lend to what

he or she is saying. The visual features garnered are used to "focus" the audio on a single subject who is speaking and to improve the quality of speech separation.

To train their joint audio-visual model, Ephrat and collaborators curated a new dataset, "AVSpeech," comprised of thousands of YouTube videos and other online video segments, such as TED Talks, how-to videos, and high-quality lectures. From AVSpeech, the researchers generated a training set of so-called "synthetic cocktail parties"—mixtures of face videos with clean speech and other speech audio tracks with background noise. To isolate speech from these videos, the user is only required to specify the face of the person in the video whose audio is to be singled out.

In multiple examples detailed in the paper, titled "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," the new method turned out superior results as compared to existing audio-only methods on pure speech mixtures, and significant improvements in delivering clear audio from mixtures containing overlapping speech and background noise in real-world scenarios. While the focus of the work is speech separation and enhancement, the team's novel method could also be applied to automatic speech recognition (ASR) and video transcription—i.e., closed captioning capabilities on streaming videos and TV. In a demonstration, the new joint audio-visual model produced more accurate captions in scenarios where two or more speakers were involved.

Surprised at first by how well their method worked, the researchers are excited about its future potential.

"We haven't seen speech separation done 'in-the-wild' at such quality before. This is why we see an exciting future for this technology," notes Ephrat. "There is more work needed before this technology lands in

consumer hands, but with the promising preliminary results that we've shown, we can certainly see it supporting a range of applications in the future, like video captioning, video conferencing, and even improved hearing aids if such devices could be combined with cameras."

The researchers are currently exploring opportunities for incorporating it into various Google products.

Provided by Association for Computing Machinery