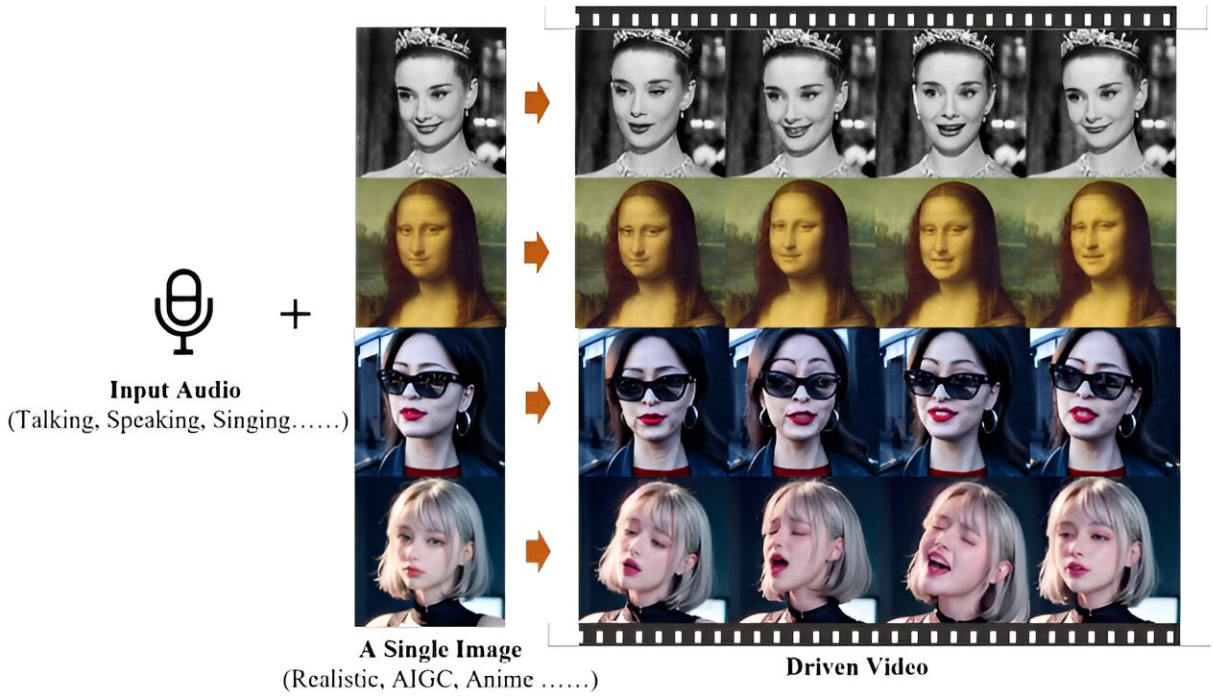
AI system can convert voice track to video of a person speaking using a still image

March 1 2024, by Bob Yirka



We proposed EMO, an expressive audio-driven portrait-video generation framework. Input a single reference image and the vocal audio, e.g. talking and singing, our method can generate vocal avatar videos with expressive facial expressions, and various head poses, meanwhile, we can generate videos with any duration depending on the length of input audio. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2402.17485

A small team of artificial intelligence researchers at the Institute for

Intelligent Computing, Alibaba Group, demonstrates, via videos they created, a new AI app that can accept a single photograph of a person's face and a soundtrack of someone speaking or singing and use them to create an animated version of the person speaking or singing the voice track. The group has published a paper describing their work on the *arXiv* preprint server.

Prior researchers have demonstrated AI applications that can process a photograph of a face and use it to create a semi-animated version. In this new effort, the team at Alibaba has taken this a step further by adding sound. And perhaps, just as importantly, they have done so without the use of 3D models or even facial landmarks. Instead, the team has used diffusion modeling based on training an AI on large datasets of audio or video files. In this instance, the team used approximately 250 hours of such data to create their app, which they call Emote Portrait Alive (EMO).

By directly converting the audio waveform into video frames, the researchers created an application that captures subtle human facial gestures, quirks of speech and other characteristics that identify an animated image of a face as human-like. The videos faithfully recreate the likely mouth shapes used to form words and sentences, along with expressions typically associated with them.

The team has posted several videos demonstrating the strikingly accurate performances they generated, claiming that they outperform other applications regarding realism and expressiveness. They also note that the finished video length is determined by the length of the original audio track. In the videos, the original picture is shown alongside that person speaking or singing in the voice of the person who was recorded on the original audio track.

The team concludes by acknowledging that use of such an application

will need to be restricted or monitored to prevent unethical use of such technology.

More information: Linrui Tian et al, EMO: Emote Portrait Alive—Generating Expressive Portrait Videos with Audio2Video Diffusion Model under Weak Conditions, *arXiv* (2024). DOI: 10.48550/arxiv.2402.17485

EMO: humanaigc.github.io/emote-portrait-alive/

Citation: AI system can convert voice track to video of a person speaking using a still image (2024, March 1) retrieved 9 May 2024 from https://techxplore.com/news/2024-03-ai-voice-track-video-person.html