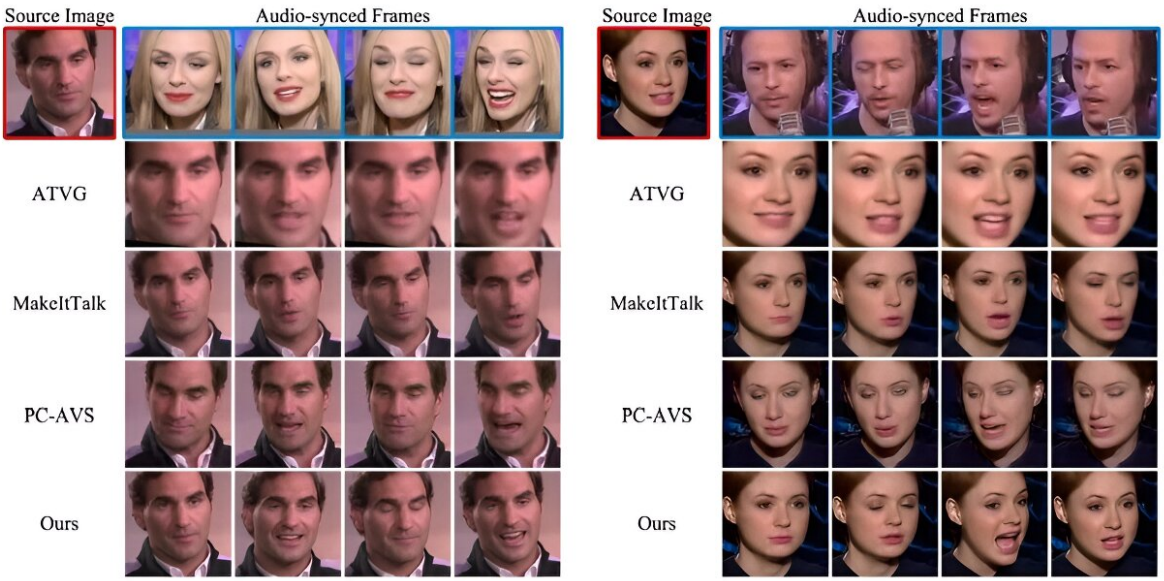
Creating realistic 'talking heads' with an AI-powered program

November 16 2023



Comparisons of DIRFA with state-of-the-art audio-driven talking face generation approaches. Credit: Nanyang Technological University

A team of researchers led by Assoc Prof Lu Shijian from the NTU School of Computer Science and Engineering has developed a computer program that creates realistic videos that reflect the facial expressions and head movements of the person speaking, only requiring an audio clip and a face photo.

DIverse yet Realistic Facial Animations, or DIRFA, is an artificial intelligence-based program that takes audio and a photo and produces a 3D video showing the person demonstrating realistic and consistent facial animations synchronized with the spoken audio. The NTU-developed program improves on existing approaches, which struggle with pose variations and emotional control.

To accomplish this, the team trained DIRFA on more than 1 million audiovisual clips from more than 6,000 people derived from an open-source database to predict cues from speech and associate them with facial expressions and head movements.

The researchers said DIRFA could lead to new applications across various industries and domains, including health care, as it could enable more sophisticated and realistic virtual assistants and chatbots, improving user experiences. It could also serve as a powerful tool for individuals with speech or facial disabilities, helping them to convey their thoughts and emotions through expressive avatars or digital representations, enhancing their ability to communicate.

Corresponding author Associate Professor Lu Shijian, from the School of Computer Science and Engineering (SCSE) at NTU Singapore, who led the study, said, "The impact of our study could be profound and far-reaching, as it revolutionizes the realm of multimedia communication by enabling the creation of highly realistic videos of individuals speaking, combining techniques such as AI and machine learning.

"Our program also builds on previous studies and represents an advancement in the technology, as videos created with our program are complete with accurate lip movements, vivid facial expressions and natural head poses, using only their audio recordings and static images."

First author Dr. Wu Rongliang, a Ph.D. graduate from NTU's SCSE,

said, "Speech exhibits a multitude of variations. Individuals pronounce the same words differently in diverse contexts, encompassing variations in duration, amplitude, tone, and more. Furthermore, beyond its linguistic content, speech conveys rich information about the speaker's emotional state and identity factors such as gender, age, ethnicity, and even personality traits.

"Our approach represents a pioneering effort in enhancing performance from the perspective of audio representation learning in AI and machine learning." Dr. Wu is a Research Scientist at the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore.

The findings were published in the journal *Pattern Recognition*.

Speaking volumes: Turning audio into action with animated accuracy

The researchers say that creating lifelike facial expressions driven by audio poses a complex challenge. For a given audio signal, there can be numerous possible facial expressions that would make sense, and these possibilities can multiply when dealing with a sequence of audio signals over time.

Since audio typically has strong associations with lip movements but weaker connections with facial expressions and head positions, the team aimed to create talking faces that exhibit precise lip synchronization, rich facial expressions, and natural head movements corresponding to the provided audio.

To address this, the team first designed their AI model, DIRFA, to capture the intricate relationships between audio signals and facial animations. Assoc Prof Lu added, "Specifically, DIRFA modeled the

likelihood of a facial animation, such as a raised eyebrow or wrinkled nose, based on the input audio. This modeling enabled the program to transform the audio input into diverse yet highly lifelike sequences of facial animations to guide the generation of talking faces.

"Extensive experiments show that DIRFA can generate talking faces with accurate lip movements, vivid facial expressions and natural head poses. However, we are working to improve the program's interface, allowing certain outputs to be controlled. For example, DIRFA does not allow users to adjust a certain expression, such as changing a frown to a smile."

In addition to adding more options and improvements to DIRFA's interface, the NTU researchers will be finetuning its facial expressions with a wider range of datasets that include more varied facial expressions and voice audio clips.

More information: Rongliang Wu et al, Audio-driven talking face generation with diverse yet realistic facial animations, *Pattern Recognition* (2023). DOI: 10.1016/j.patcog.2023.109865. On *arXiv*: DOI: 10.48550/arxiv.2304.08945

Provided by Nanyang Technological University