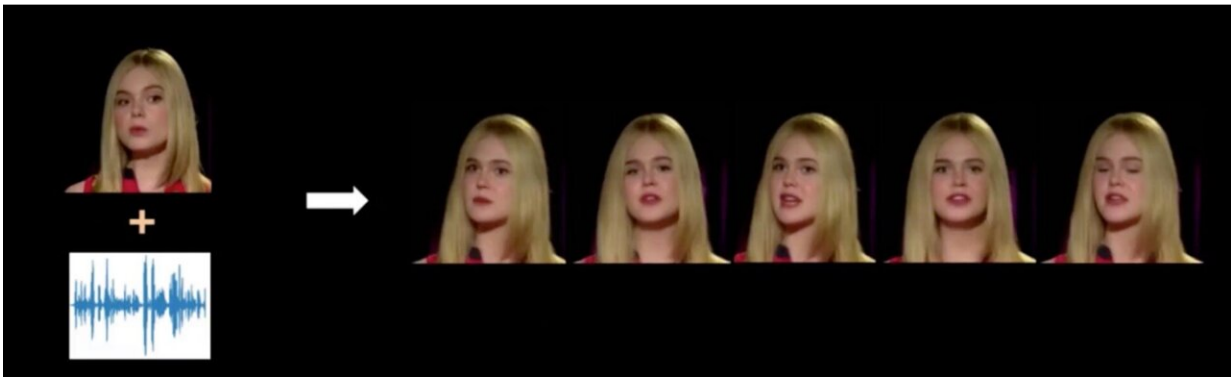# A model that can create realistic animations of talking faces

January 28 2022, by Ingrid Fadelli

Credit: Biswas et al.

In recent years, computer-generated animations of animals and humans have become increasingly detailed and realistic. Nonetheless, producing convincing animations of a character's face as it's talking remains a key challenge, as it typically entails the successful combination of a series of different audio and video elements.

A team of computer scientists at TCS Research in India has recently created a new model that can produce highly realistic talking face animations that integrate audio recordings with a character's head motions. This model, introduced in a paper presented at ICVGIP 2021, the twelfth Indian Conference on Computer Vision, Graphics and Image Processing, could be used to create more convincing virtual avatars,

digital assistants, and animated movies.

"For a pleasant viewing experience, the perception of realism is of utmost importance, and despite recent research advances, the generation of a realistic talking face remains a challenging research problem," Brojeshwar Bhowmick, one of the researchers who carried out the study, told TechXplore. "Alongside accurate lip synchronization, realistic talking face animation requires other attributes of realism such as natural eye blinks, head motions and preserving identity information of arbitrary target faces."

Most existing speech-driven methods for generating face animations focus on ensuring a good synchronization between lip movements and recorded speech, preserving a character's identity and ensuring that it occasionally blinks its eyes. A few of these methods also tried to generate convincing head movements, primarily by emulating those performed by human speakers in a short training video.

"These methods derive the head's [motion](#) from the driving video, which can be uncorrelated with the current speech content and hence appear unrealistic for the animation of long speeches," Bhowmick said. "In general, head motion is largely dependent upon the prosodic information of the speech at a current time window."

Past studies have found that there is a strong correlation between the head movements performed by human speakers and both the pitch and amplitude of their voice. These findings inspired Bhowmick and his colleagues to create a new method that can produce head motions for face animations that reflect a character's voice and what he/she is saying.

In [one of their previous papers](#), the researchers presented a generative adversarial network (GAN)-based architecture that could generate convincing animations of faces talking. While this technique was

promising, it could only produce animations in which the head of speakers did not move.

"We now developed a complete speech-driven realistic facial animation pipeline that generates talking face videos with accurate lip-sync, natural eye-blinks and realistic head motion, by devising a hierarchical approach for disentangled learning of motion and texture," Bhowmick said. "We learn speech-induced motion on facial landmarks, and use the landmarks to generate the texture of the animation video frames."

The new generative model created by Bhowmick and his colleagues can effectively generate speech-driven and realistic head movements for animated talking faces, which are strongly correlated with a speaker's vocal characteristics and what he/she is saying. Just like the technique they created in the past, this new model is based on GANs, a class of machine learning algorithms that has been found to be highly promising for generating artificial content.

The model can identify what a speaker is talking about and his/her voice's intonation during specific time windows. Subsequently, it uses this information to produce matching and correlated head movements.

"Our method is fundamentally different from state-of-the-art methods that focus on generating person-specific talking style from the target subject's sample driving video," Bhowmick said. "Given that the relationship between the audio and head motion is not unique, our attention mechanism tries to learn the importance of local audio features to the local head motion keeping the prediction smooth over time, without requiring any input driving video at test time. We also use meta-learning for texture generation, as it helps to quickly adapt to unknown faces using very few images at test time."

Bhowmick and his colleagues evaluated their model on a series of

benchmark datasets, comparing its performance to that of state-of-the-art techniques developed in the past. They found that it could generate highly convincing animations with excellent lip synchronization, natural eye blinks, and speech-coherent head motions.

"Our work is a step further towards achieving realistic talking face animations that can translate into multiple real-world applications, such as digital assistants, video dubbing or telepresence," Bhowmick added. "In our next studies, we plan to integrate realistic facial expressions and emotions alongside lip sync, eye blinks and speech-coherent head motion."

More information: Dipanjan Das et al, Speech-driven facial animation using cascaded GANs for learning of motion and texture. *European Conference on Computer Vision* (2020). www.ecva.net/papers/eccv_2020/ … papers/123750409.pdf

© 2022 Science X Network