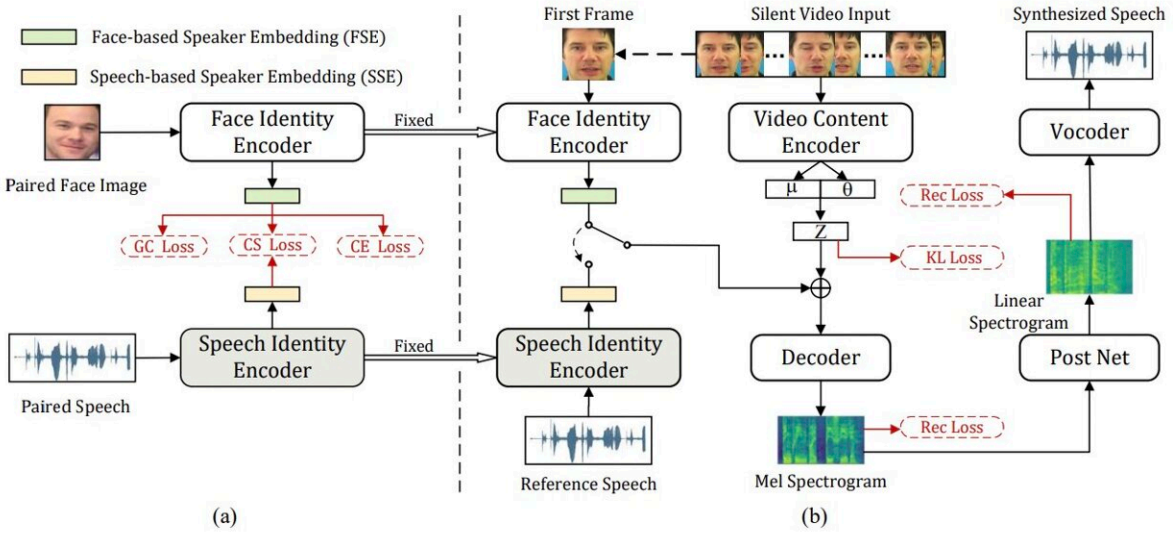
A model that can create synthetic speech that matches a speaker's lip movements

June 13 2023, by Ingrid Fadelli



The overall flowchart of the team's proposed method. Credit: Sheng, Ai and Ling

Machine learning models can help to solve several real-world problems faster and more efficiently. One among these problems involves synthesizing speech for both animated characters and human speakers based on the movements of their lips.

To tackle this task, known as lip-to-speech (Lip2Speech) synthesis, machine learning models essentially learn to predict what spoken words would result from specific sequences of face and lip movements.

Automating Lip2Speech synthesis could be useful for numerous use cases, for instance helping patients who cannot produce speech sounds to communicate with others, adding sound to silent movies, restoring speech in noisy or damaged videos, and even for determining what potential criminals are saying in voice-less CCTV footage.

While some machine learnings for Lip2Speech applications achieved promising results, most of these models perform poorly in real-time and are not trained using so-called zero-shot learning approaches. Zero-shot learning essentially means that a pre-trained model can effectively make predictions related to data classes that it did not encounter during training.

Researchers at the University of Science and Technology of China recently developed a new model for Lip2Speech synthesis that can produce personalized synthesized speech in zero-shot conditions. This approach, introduced in a paper published on the *arXiv* pre-print server, is based on a variational autoencoder, a generative model partly based on neural networks that encode and decode data.

To effectively tackle Lip2Speech tasks in zero-shot conditions, machine learning models would typically need to extract additional information about speakers from reliable video recordings of them speaking. However, if only silent or unintelligible videos of their face speaking are available, this information cannot be accessed. The model created by this team of researchers could circumvent this issue, by generating speech that matches the appearance and identity of a given speaker without requiring recordings of the speaker's actual speech.

"We propose a zero-shot personalized Lip2Speech synthesis method, in which face images control speaker identities," Zheng-Yan Sheng, Yang Ai, and Zhen-Hua Ling wrote in their paper. "A variational autoencoder is adopted to disentangle the speaker identity and linguistic content

representations, which enables speaker embeddings to control the voice characteristics of synthetic speech for unseen speakers. Furthermore, we propose associated cross-modal representation learning to promote the ability of face-based speaker embeddings (FSE) on voice control."

Sheng, Ai and Ling evaluated their model in a series of tests and found that it performed remarkably well, producing synthesized speech that matched both a speaker's lip movements and their age, gender and overall appearance. In the future, the new model could be used to create tools for a wide range of applications, including assistive applications for people with speech impairments, video editing tools and software to aid police investigations.

"Extensive experiments verify the effectiveness of the proposed method whose synthetic utterances are more natural and matching with the personality of input video than the compared methods," Sheng, Ai and Ling said. "To our best knowledge, this paper makes the first attempt on zero-shot personalized Lip2Speech synthesis with a face image rather than reference audio to control voice characteristics."

More information: Zheng-Yan Sheng et al, Zero-shot personalized lip-to-speech synthesis with face image based voice control, *arXiv* (2023). DOI: 10.48550/arxiv.2305.14359