ExpressiveTalk: Emotionally and Stylistically Enhanced Talking Face Generation

Thiranja Jayasundara, Mathis Patigny, Serden-Yilmaz Kose, Corentin Nicodème, Ahmed Shamsia,
and Mohamed Mohamed
University of Oulu, Finland

Research Plan

Several systems have been introduced over the years to create a speaking face from a picture or a portrait, while providing an audio input to speak in a short video. However, these systems were made with the intention of preserving the emotion or the speaker's style. For example, *StyleTalk* (Ma et al., 2023) was created to capture and transfer the speaker's style, including tempo, rhythm, and pace, from an arbitrary reference speaking video and then drive the one-shot portrait to speak with a reference speaking style and another piece of audio. Such a system lacked the concept of direct control over the speaker's emotions. On the other hand, *EmoGen* (Goyal et al., 2023) allowed the addition of emotions from a small emotions set, such as happiness, anger, and sadness, to the reference portrait and an audio clip. The system struggled to preserve the speaker's subtle speaking style while displaying unrealistic speaking faces, which is not typically the case in natural conversations. In Summary, *StyleTalk* provided a speaking style with no controlled emotions, while *EmoGen* provided controlled emotions without a speaking style. Therefore, *ExpressiveTalk* is proposed as a hybrid system to bridge the gap by creating a talking-face system that provides a more realistic speaking face by combining the emotions and style of the speaker into a single system.

ExpressiveTalk is a web interface application that takes a single portrait as input, along with an audio clip and either a chosen emotion label or a short style reference video. The system then fuses these inputs using a lightweight emotion encoder and a style-aware transformer, supported by a lip-sync backbone for precise mouth movements. This design enables the rendering of emotions such as happiness, sadness, or subtle blends, while still capturing the tempo and intensity of speech from a reference clip, when available. By uniting the controllability of *EmoGen* with the naturalism of *StyleTalk*, *ExpressiveTalk* aims to reduce the "uncanny valley" effect, where emotionally enhanced faces sometimes look unnatural, producing consistent, emotionally rich talking heads.

Reference

Goyal, S., Uppal, S., Bhagat, S., Yu, Y., Yin, Y., & Shah, R. R. (2023). Emotionally enhanced talking face generation. https://arxiv.org/abs/2303.11548

Ma, Y., Wang, S., Hu, Z., Fan, C., Lv, T., Ding, Y., Deng, Z., & Yu, X. (2023). Styletalk: One-shot talking head generation with controllable speaking styles. https://arxiv.org/abs/2301.01081