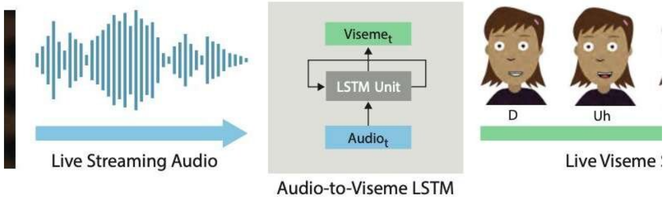
A deep learning technique to generate real-time lip sync for live 2-D animation

11 November 2019, by Ingrid Fadelli



Real-Time Lip Sync. Our deep learning approach uses an LSTM to convert live streaming audio to discrete visemes for 2D characters. Credit: Aneja & Li.

Live 2-D animation is a fairly new and powerful form of communication that allows human performers to control cartoon characters in real time while interacting and improvising with other actors or members of an audience. Recent examples include Stephen Colbert interviewing cartoon guests on *The Late Show*, Homer answering live phone-in questions from viewers during a segment of *The Simpsons*, Archer talking to a live audience at ComicCon, and the stars of Disney's *Star vs. The Forces of Evil* and *My Little Pony* hosting live chat sessions with fans via YouTube or Facebook Live.

Producing realistic and effective live 2-D animations requires the use of interactive systems that can automatically transform human performances into animations in real time. A key aspect of these systems is attaining a good lip sync, which essentially means that the mouths of animated characters move appropriately when speaking, mimicking the movements observed in the mouths of performers.

Good lip syncing can make live 2-D animation more convincing and powerful, allowing animated characters to embody the performance more realistically. Conversely, poor lip syncing typically breaks the illusion of characters as live participants in a performance or dialog.

In a paper recently prepublished on *arXiv*, two researchers at Adobe Research and the University of Washington introduced a deep learning-based interactive system that automatically generates live lip sync for layered 2-D animated characters. The system they developed uses a long short-term memory (LSTM) model, a recurrent neural network (RNN) architecture often applied to tasks that involve classifying or processing data, as well as making predictions.

"Since speech is the dominant component of almost every live animation, we believe the most critical problem to address in this domain is live lip sync, which entails transforming an actor's speech into corresponding mouth movements (i.e., viseme sequence) in the animated character. In this work, we focus on creating high-quality lip sync for live 2-D animation," Wilmot Li and Deepali Aneja, the two researchers who carried out the research, told TechXplore via email.

Li is a principal scientist at Adobe Research with a Ph.D. in computer science who has been conducting extensive research focusing on topics at the intersection of computer graphics and human-computer interaction. Aneja, on the other hand, is currently completing a Ph.D. in computer science at the University of Washington, where she is part of the Graphics and Imaging Lab.

The system developed by Li and Aneja uses a simple LSTM model to convert streaming audio input into a corresponding viseme sequence at 24 frames per second, with less than 200 milliseconds latency. In other words, their system allows an animated character's lips to move in a similar way to those of a human user speaking in real time, with less than 200 milliseconds of delay between the voice and the lip movement.

"In this work, we make two contributions—identifying the appropriate feature representation and network configuration to achieve state-of-the-art results for live 2-D lip sync and devising a new augmentation method for collecting training data for the model," Li and Aneja explained.

"For hand-authoring lip sync, professional animators make stylistic decisions about the specific choice of visemes and the timing and number of transitions. As a result, training a single 'general-purpose' model is unlikely to be sufficient for most applications," Li and Aneja said. Furthermore, obtaining labeled lip sync data to train deep learning models can be both expensive and time-consuming. Professional animators can spend five to seven hours of work per minute of speech to hand-author viseme sequences. Aware of these limitations, Li and Aneja developed a method that can generate training data faster and more effectively.

To train their LSTM model more effectively, Li and Aneja introduced a new technique that augments hand-authored training data using audio time warping. This data augmentation procedure achieved good lip syncing even when training their model on a small labeled dataset.

To evaluate the effectiveness of their interactive system in producing lip sync in real time, the researchers asked human viewers to rate the quality of live animations powered by their model with those produced using commercial 2-D animation tools. They found that most viewers preferred the lip sync generated by their approach over that produced by other techniques.

"We also investigated the trade-off between lip sync quality and the amount of training data, and we found that our data augmentation method significantly improves the output of the model," Li and Aneja said. "In general, we can produce reasonable results with just 15 minutes of hand-authored lip sync data."

Interestingly, the researchers found that their LSTM model can acquire different lip sync styles based on the data it is trained on, while also generalizing well across a broad range of speakers. Impressed by the encouraging results achieved by the model, Adobe decided to integrate a version of it within its Adobe Character Animator software, released in the fall of 2018.

"Accurate, low-latency lip sync is important for almost all live animation settings, and our human judgment experiments show that our technique improves on existing state-of-the-art 2-D lip sync engines, most of which require offline processing," Li and Aneja said. Thus, the researchers believe that their work has immediate practical implications for both live and non-live 2-D animation production. The researchers are not aware of previous 2-D lip sync work with similarly comprehensive comparisons against commercial tools.

In their recent study, Li and Aneja were able to address some of the key technical challenges associated with the development of techniques for live 2-D animation. First, they demonstrated a new method to encode artistic rules for 2-D lip sync using RNNs, which could be further enhanced in the future.

The researchers believe there are many more opportunities to apply modern machine learning techniques to improve 2-D animation workflows. "Thus far, one challenge has been the lack of training data, which is expensive to collect. However, as we show in this work, there may be ways to leverage structured data and automatic editing algorithms (e.g., dynamic time warping) to maximize the utility of hand-crafted animation data," Li and Aneja said.

Although the data augmentation strategy proposed by the researchers can significantly reduce training data requirements for models designed to produce real-time lip sync, hand-animating enough lip sync content to train new models still requires considerable work and effort. According to Li and Aneja, however, retraining an entire model from scratch for every new lip sync style it encounters may be unnecessary.

The researchers are interested in exploring fine-tuning strategies that could allow animators to adapt the model to different styles with a much smaller amount of user input. "A related idea is to directly learn a lip sync model that explicitly includes tunable stylistic parameters. While this may require a much larger training dataset, the potential benefit is a model that is general enough to support a range of lip sync styles without additional training," the researchers said.

Interestingly, in their experiments, the researchers observed that the simple cross-entropy loss they used to train their model did not accurately reflect the most relevant perceptual differences between lip sync sequences. More specifically, they found that certain discrepancies (e.g., missing a transition or replacing a closed mouth viseme with an open mouth viseme) are much more obvious than others. "We think that designing or learning a perceptually based loss in future research may lead to improvements in the resulting model," Li and Aneja said.

More information: Real-time lip sync for live 2-D animation. arXiv:1910.08685 [cs.GR]. arxiv.org/abs/1910.08685

github.com/deepalianeja/CharacterLipSync

APA citation: A deep learning technique to generate real-time lip sync for live 2-D animation (2019, November 11) retrieved 21 January 2022 from https://techxplore.com/news/2019-11-deep-technique-real-time-lip-sync.html