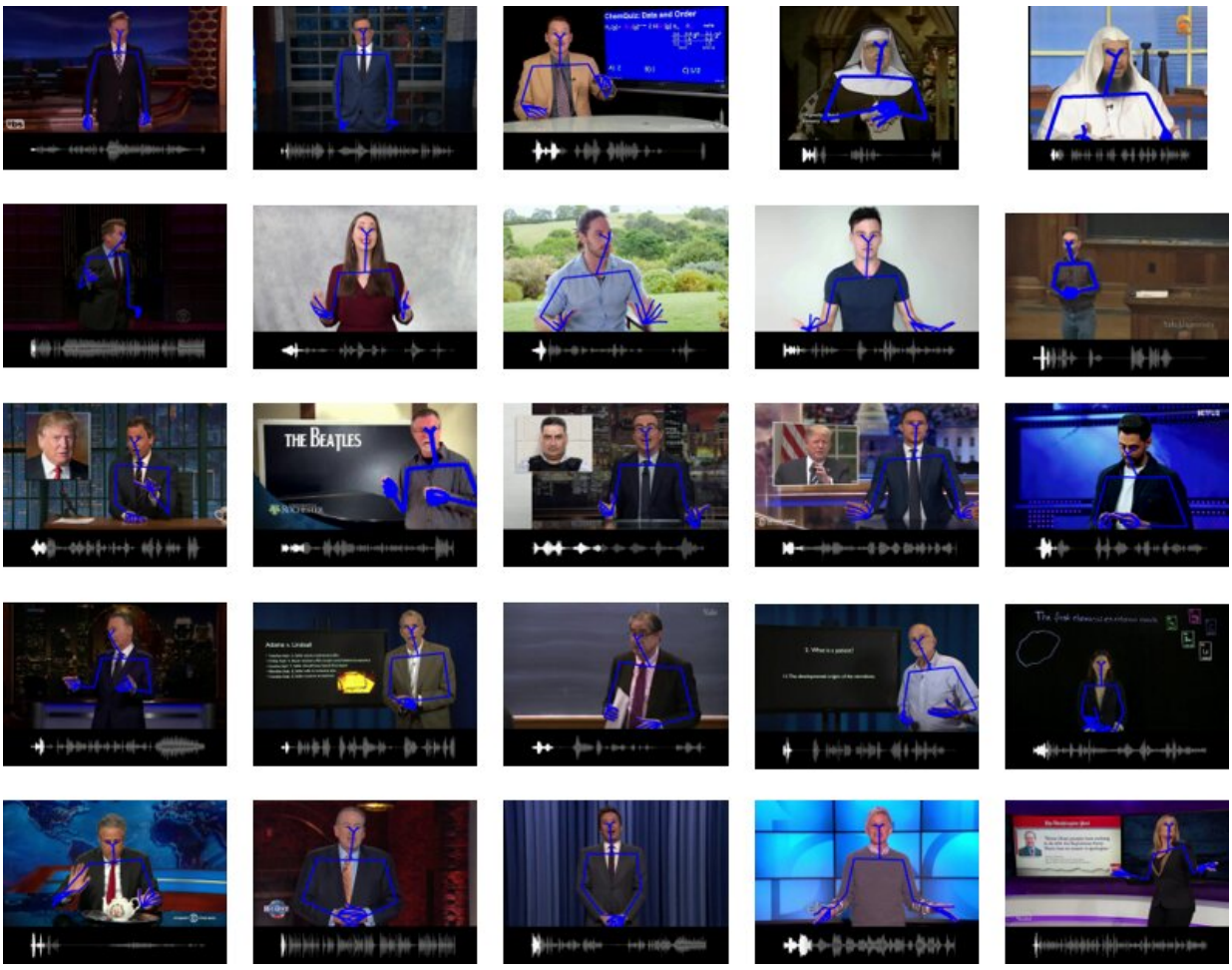
Mix-StAGE: A model that can generate gestures to accompany a virtual agent's speech

August 13 2020, by Ingrid Fadelli



An overview of PATS, the dataset for studying gesture styles compiled by the researchers. Credit: Ahuja et al.

Virtual assistants and robots are becoming increasingly sophisticated, interactive and human-like. To fully replicate human communication, however, artificial intelligence (AI) agents should not only be able to determine what users are saying and produce adequate responses, they should also mimic humans in the way they speak.

Researchers at Carnegie Mellon University (CMU) have recently carried out a study aimed at improving how virtual assistants and robots communicate with humans by generating natural gestures to accompany their speech. Their paper, pre-published on arXiv and set to be presented at the European Conference on Computer Vision (ECCV) 2020, introduces Mix-StAGE, a new model that can produce different styles of co-speech gestures that best match the voice of a speaker and what he/she is saying.

"Imagine a situation where you are communicating with a friend in a virtual space through a virtual reality headset," Chaitanya Ahuja, one of the researchers who carried out the study, told TechXplore. "The headset is only able to hear your voice, but not able to see your hand gestures. The goal of our model is to predict the hand gestures accompanying the speech."

Humans typically have unique ways of gesturing as they communicate with others. Ahuja and his colleagues wanted to create a co-speech gesture generation model that takes these individual differences into account, producing gestures that are aligned with the voice and personality of speakers.

"The key idea behind Mix-StAGE is to learn a common gesture space for many different styles of gestures," Ahuja said. "This gesture space consists of all the possible gestures, which are grouped by style. The second half of Mix-StAGE learns how to predict gestures in any given style while being synchronized with the input speech signal, a process

that is known as style transfer."

Mix-StAGE was trained to produce effective gestures for multiple speakers, learning the unique style characteristics of each speaker and producing gestures that match these characteristics. In addition, the model can generate gestures in one speaker's style for another speaker's speech. For instance, it could generate gestures that match what speaker A is saying in the gestural style typically used by speaker B.

"We were able to teach a single model (i.e., with less memory involved) to represent many gestural styles, unlike previous approaches that required a separate model for each style," Ahuja explained. "Our model takes advantage of similarities between gestural styles, while at the same time remembering what is unique about each person (i.e., each style)."

In initial tests, the model developed by Ahuja and his colleagues performed remarkably well, producing realistic and effective gestures in different styles. Moreover, the researchers found that as they increased the number of speakers used to train Mix-StAGE, its gesture generation accuracy significantly improved. In the future, the model could help to enhance the ways in which virtual assistants and robots communicate with humans.

To train Mix-StAGE, the researchers compiled a dataset called Pose-Audio-Transcript-Style (PATS), containing audio recordings of 25 different people speaking, for a total of over 250 hours, with matching gestures. This dataset could soon be used by other research teams to train other gesture generation models.

"In our current research, we focus on the nonverbal part of speech (e.g., prosody) when generating gestures," Ahuja said. "We are excited about the next step, where we will include also the verbal part of speech (i.e., language) as another input. The hypothesis is that language will help with

specific types of gestures, such as iconic or metaphorical gestures, where the meaning of the spoken words may be the most important."

More information: Style transfer for co-speech gesture animation: a multi-speaker conditional-mixture approach. arXiv:2007.12553 [cs.CV]. arxiv.org/abs/2007.12553