Unifying Speech and Gesture Synthesis

Originally published at https://www.unite.ai/unifying-speech-and-gesture-synthesis/

When I came back to Britain from some years in Southern Italy, it took quite a while to stop gesticulating while I talked. In the UK, supporting your speech with bold hand movements just makes you look over-caffeinated; in Italy, as someone learning the language, it actually helped me to be understood. Even now, on the more rare occasions that I speak Italian, the 'wild hands' come back into service. It's almost impossible to speak Italian without moving.

In recent years, gesture-supported communication in Italian and Jewish culture has come to public attention as more than just a trope from the work of Martin Scorsese and early Woody Allen movies. In 2013 the New York Times compiled a short video history of Italian hand gestures; academia is beginning to study racial propensities for hand-gesturing, rather than dismissing the subject as a stereotype; and new emojis from the Unicode Consortium are closing the gesture shortfall that comes with purely digital, text-based communication.

A Unified Approach to Speech and Gesticulation

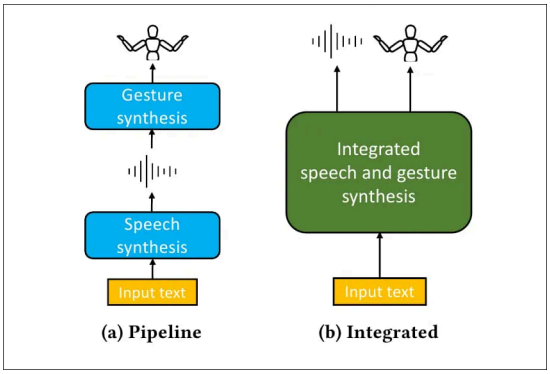
Now, new research from the Department of Speech, Music and Hearing at Sweden's KTH Royal Institute of Technology is seeking to combine speech and gesture recognition into a unified, multi-modal system that could potentially increase our understanding of speech-based communication by using body language as an integrated adjunct to speech, rather than a parallel field of study.

CLICK TO VIEW ANIMATED GIF

Visuals from the test page of the Swedish speech/gesture project. Source: https://swatsw.github.io/isg_icmi21/

The research proposes a new model called Integrated Speech and Gesture (ISG) synthesis, and brings together a number of state-of-the-art neural models from speech and gesture research.

The new approach abandons the linear pipeline model (where gesture information is derived sequentially from speech as a secondary processing stage) for a more integrated approach, which rates equally with existing systems according to end users, and which achieves faster synthesis time and reduced parameter count.



Linear vs. integrated approaches. Source: https://arxiv.org/pdf/2108.11436.pdf

The new multimodal system incorporates a spontaneous text-to-speech synthesizer and an audio-speech-driven gesture generator, both trained on the existing Trinity Speech Gesture dataset. The dataset contains 244 minutes of audio and body capture of a man talking on different topics and gesticulating freely.

The work is a novel and tangential equivalent to the DurIAN project, which generates facial expressions and speech, rather than gesture and speech, and which falls more into the realm of expression recognition and synthesis.

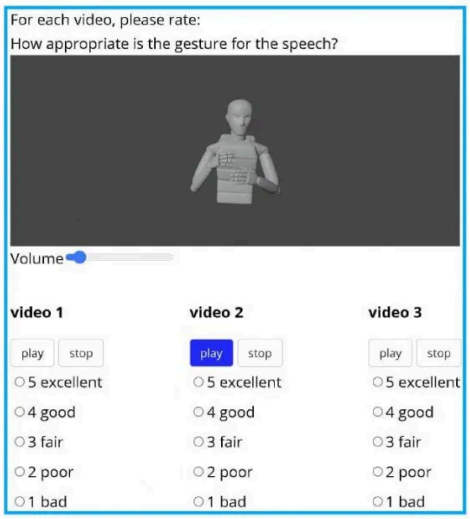
Architectures

The speech and visual (gesture) components of the project are ill-balanced in terms of data; text is sparse and gesticulation is rich and data-intensive – a challenge in terms of defining goals and metrics. Therefore the researchers evaluated the system primarily by human response to the output, rather than more obvious mechanistic approaches such as mean square error (MSE).

The two main ISG models were developed around the second iteration of Google's 2017 Tacotron end-to-end speech synthesis project, and the South Korean Glow-TTS initiative published in 2020. Tacotron utilizes an autoregressive LSTM architecture, while Glow-TTS acts in parallel via convolution operators, with faster GPU performance and without the stability issues that can attend autoregressive models.

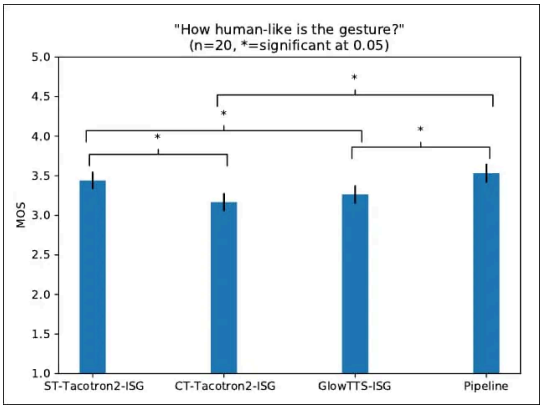
The researchers tested three effective speech/gesture systems during the project: a modified version of a multimodal speech-and-gesture-generation published in 2021 by a number of the same researchers on the new project; a dedicated and modified ISG version of the open source Tacotron 2; and a highly altered ISG version of Glow-TTS.

To evaluate the systems, the researchers created a web-based feedback environment featuring articulated 3D people speaking and moving to predefined text segments (the general look of the environment can be seen at the public project page).



The test environment.

Test subjects were asked to evaluate system performance based on speech and gesture, speech only, and gesture only. The results showed a slight improvement in the new ISG version over the older pipeline version, though the newer system operates more quickly and with reduced resources.



Asked 'How human is the gesture?', the fully integrated ISG model finishes slightly ahead of the slower pipeline model, with the Tacotron and Glow-based models further behind.

Embedded Shrug

The Tacotron2-ISG model, the most successful of the three approaches, demonstrates a level of 'subliminal' learning related to some of the most common phrases in the dataset, such as 'I don't know' – despite a lack of explicit data that would cause it to generate a shrug to accompany this phrase, the researchers found that the generator does indeed shrug.

The researchers note that the very specific nature of this novel project inevitably means a scarcity of general resources, such as dedicated datasets that incorporate speech and gesture data in a way that's suitable for training such a system. Nonetheless, and in spite of the vanguard nature of the research, they consider it a promising and little-explored avenue in speech, linguistics and gesture recognition.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

UNITE.AI

Discover More