CELLULAIRE EN COLÈRE

angry cellphone

Jeena Yin, Anirudh Mani, Joseph Gibli, Zaria Howard

concept

Machine learning has given technology a human voice. Even though models can produce human-like audio, computer generated speech has almost solely been used to serve end-users as neutral sounding voice assistants. Our project grants technology the power of free speech by training various neural networks on emotional audio recordings. We allow these models to produce emotive speech and then direct their frustration towards each other. Human spectators to their argument are faced with learned rage communicated through tools typically meant to mindlessly serve.

background

"The Quintet of the Astonished" – Bill Viola – The artist presents a video of five actors emotions unfolding in extreme slow motion so every minute detail of their cathartic expression can be explored by the viewer.

SampleRNN: An Unconditional End-to-End Neural Audio Generation Model – Soroush Mehri, et al. – This paper proposes a model for unconditional audio generation through an autoregressive multilayer perceptron and hierarchical recurrent neural network. Human evaluation of audio samples indicates SampleRNN is preferred over other models.

WaveRNN: Efficient Neural Audio Synthesis – Nal Kalchbrenner, et al. – This paper presents a new version of neural audio synthesis using a single-layer recurrent neural network with a dual soft max layer. Generated audio matches the quality of the state-of-the-art WaveNet model with four times speed.

approach

Cellulaire en Colère differs from existing literature because certain models were trained strictly on angry audio. This allows our model to learn the waveform and rhythm of speech communicated angrily. We also do not limit our model to English words in order to grant the computer freedom of expression.

citations

Viola, Bill. "The Quintet of the Astonished (excerpt)". *uploaded to vimeo.com by Urban Video Project*

Kalchbrenner, Nal; et al. "Efficient Neural Audio Synthesis". 23 Feb 2018. *arXiv.org*

Mehri, Soroush; et al. "SampleRNN: AN Unconditional End-to-End Neural Audio Generation Model". 11 Feb 2017. *uploaded to arXiv.org*

methods

Audio Recordings – Actors recorded angry monologues using high quality microphones which were then exported to high quality WAV files. Each actor recorded roughly four minutes of audio, for a total of 15 minutes. Additionally, an audiobook recording of Ellen DeGeneres was downloaded in WAV format to determine if generated audio can present a spectrum of emotions.

SampleRNN – Since SampleRNN requires a large dataset to produce high quality output, the neural network was trained on the Ellen DeGeneres audiobook. Using default parameters, the model was trained for over 48 hours.

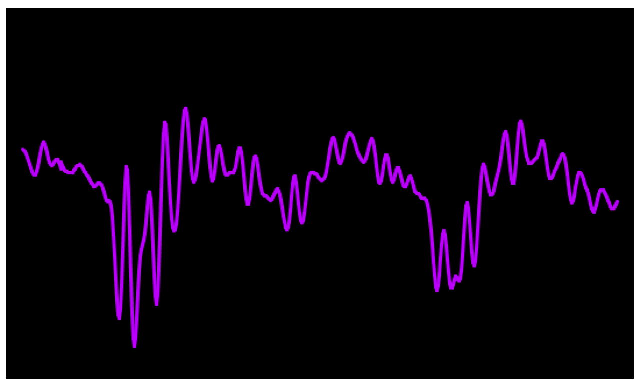
WaveRNN – We trained the model on 20 minutes of audio in two steps: once with learning rate 1e-3 and then with learning rate 1e-4. The sequential training speeds up training and produces higher quality results. In total, the model was trained for around 40 hours.

Computer Pulse – In order to visualize the audio, we created an audio waveform visualizer using Processing's Sound functionality. The Fourier transforms were converted to sine waves and presented on a black background. This conveys a feel of the computers' pulse given their generated speech.

results

The generated audio was largely successful, especially the Ellen DeGeneres SampleRNN model. SampleRNN was able to create sound that is recognizably her voice with varied pitch and sentiment. WaveRNN produced audio with exclusively ominous sentiment. The audio was more noisy with less consistent results. Although WaveRNN was trained on a high pitched female voice the generated audio had rather low pitch.

Both models freely formed babbling sounds that seem near human-like. However, some post processing was required to select the best generated samples from either model and compose them into a conversation.



Screenshot of computer pulse visualizer. Amplitude of the sine wave correspond to waveform and intensity of generated audio.