Executive summary

Quick and accurate identification of emotion leads to more effective interpersonal interactions. Using machine learning, we develop a dense neural network model that matches the accuracy of the top quartile of human reviewers. Classification of emotion based on audio is a challenging problem. While statistical tests of extracted features suggest that different emotions have different population means, plotted distributions revealed near complete overlap among every emotion meaning that any one measure had very little predictive power.

Combining RAVDESS and CREMA-D, the study data consists of over 8,000 short audio clips generated by more than 100 different actors expressing happiness, sadness, fear, disgust, anger and a neutral control. Both of the source datasets were generated by government-funded research teams associated with major universities.

Using machine learning, how can we correctly identify emotion in spoken language

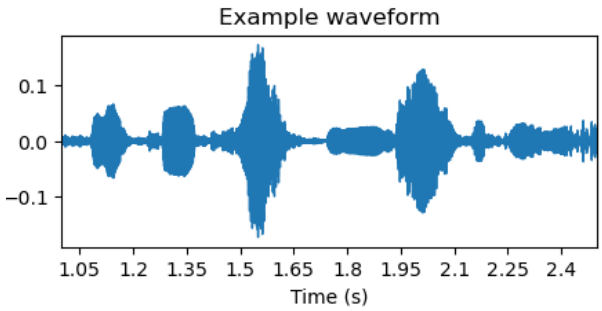
Emotion is woven throughout human experience and communication. Identifying it quickly and correctly creates value for the speaker and listener. Consider a call center where agents handle high volumes of requests and complaints each hour. Best practice is to triage angry or disgusted callers for special treatment before they post a negative review or report the company to the better business bureau. While agents are naturally able to identify some of these emotional states, their identification ability declines over the course of the work period as they become tired, distracted and less empathetic. Machine learning tools that aid in emotion identification can improve the triage flow and reduce negative outcomes.

Consider the couple that falls into an argument when one misperceives excitement for anger. Most people do not set out to miscommunicate, but over the course of a relationship, there are times of distraction and exhaustion. Tools that provide feedback to the speaker on what emotion they are expressing could help them modulate tone and thereby avoid friction.

With the proliferation of digital communication tools, audio data is ever more available. Despite the increasing use of video/audio combinations, substantial amounts of communication are still audio only. Extracting useful emotional insights from this bountiful datatype creates value for both speakers and listeners.

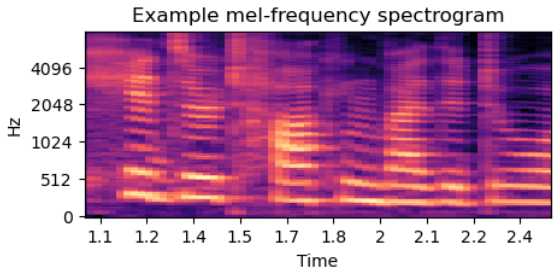
What is sound?

Before addressing the dataset, we briefly examine the nature of sound and its digital representations. Sounds are most typically experienced as variations in air pressure. These fluctuations cause small bones in the ear canal to vibrate, sending a signal to the brain. The variations in pressure and the vibration they cause can be described as a wave, with the key characteristics being frequency and amplitude. Frequency is measured in Hertz as the number of high/low pressure cycles that take place in a second. The human ear is attuned to hear frequencies ranging from 20Hz to over 20,000 Hz. Amplitude is the intensity of pressure variation; humans interpret that as loudness and measure it in decibels. Ultimately, sound can be described as a waveform such as the one shown here, with time as the x-axis and relative change in air pressure as the y-axis.



When sound is digitized, measurements of the amplitude are taken at regular intervals. Since the human ear generally can hear sounds with frequencies up to 20 kHz, moderate quality audio is sampled at 22kHz.

For analysis, audio is generally split into its component frequencies using a fourier transformation. Time, frequency and amplitude create three dimensions for analyzing sound. This is often visualized as a spectrogram (shown below), where time is the x-axis, frequency is the y-axis and amplitude is represented by color. Using the waveform or the spectrogram, various summary statistics can be calculated, either over small increments of time or over longer durations.

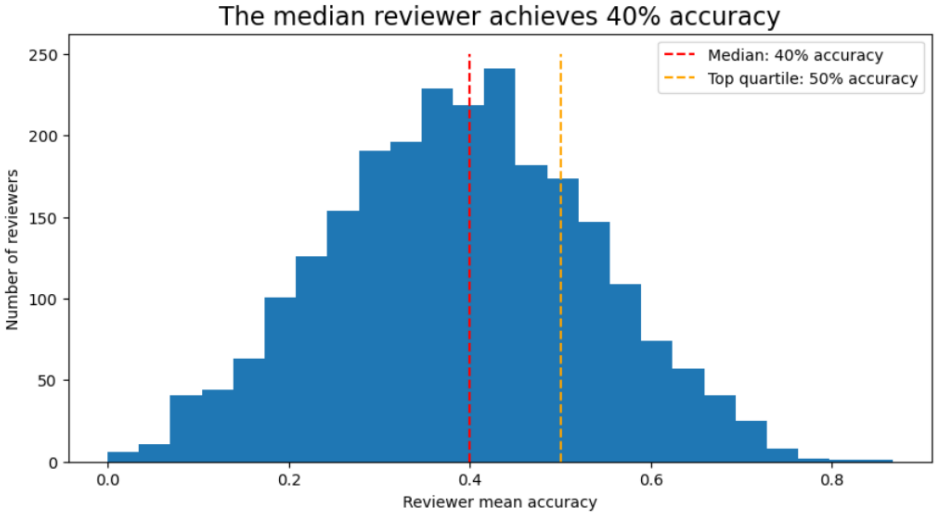


Labeled audio files from government-funded research sets are used for analysis

Our analysis used labeled audio observations from the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). These datasets both consist of actors reading short, simple sentences with no intrinsic emotional value in the written language. Some examples are “It’s eleven o’clock” and “Dogs are sitting by the door”. Actors each sentence with all six emotional classes. Afterwards, the emotional content was checked and validated with blind classification outside reviewers. CREMA-D consists of ~7,500 audio files spread across 91 actors, while RAVDESS has over 1,000 audio files spread across 24 actors. Further examination of the datasets is included in the notebooks. A dataset citation is included at the end of this report.

Performance baseline

CREMA-D includes over 70,000 assessments of its audio observations by almost 2,500 human reviewers. We use the accuracy of these reviewers as our baseline. The median reviewer achieved an accuracy of 40% and the top quartile of reviewers achieved 50% accuracy.



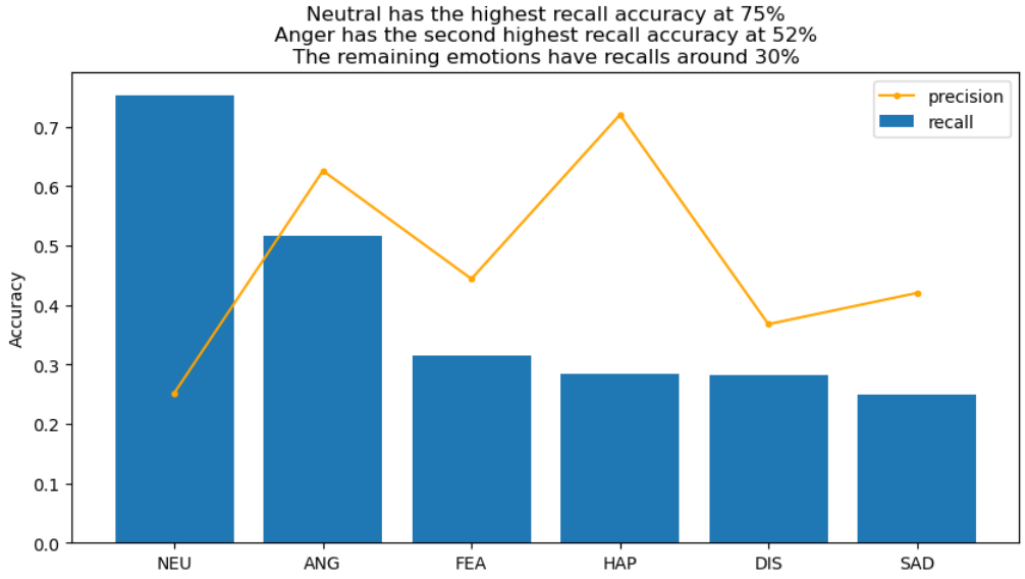
On first viewing, this accuracy seems lower than would be expected. As social animals, humans have evolved to perceive emotional states. However, assessing this particular dataset comes with unique challenges.

First, we do not know who the speakers are. Often, humans' most frequent and most important interactions involve people whom we have met before and with whom they have some prior relationship. Memories of previous interactions create an understanding of the speaker's motivations and prior emotional states. This understanding provides significant contributions to the task of identifying emotion. In this dataset, there is no context.

Second, over the course of human evolution, most communication took place at close range where the speaker's body language and facial expressions could be observed. Given the historical ubiquity of this visual information, humans tend to rely on it for emotional classification. The absence of accompanying visual data in this dataset increases the difficulty of correctly identifying emotion.

Lastly, the words in most verbal communication contain some emotional content. For this dataset, sentences were specifically chosen to have meanings devoid of emotional content.

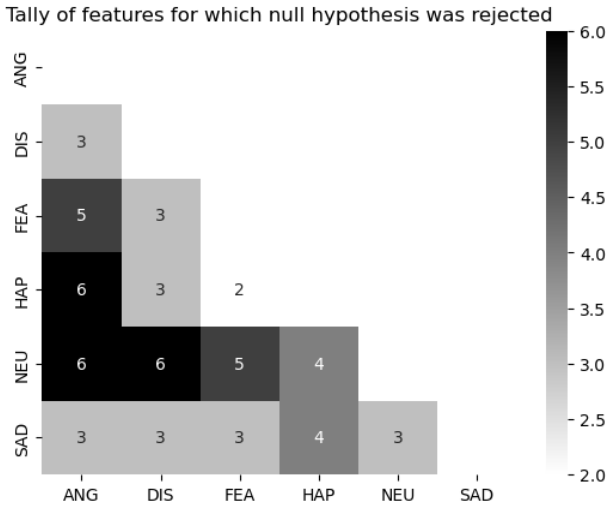
Diving deeper into human reviewer performance, it is worth noting the heavy-handedness with which people used the neutral classification. This resulted in excellent recall accuracy for the neutral category, but very low precision. Anger seemed to be the easiest to identify as exhibited by it having both a high recall and high precision score. Anger suggest an imminent threat, so evolution may have prioritized recognition of this emotional state.



Exploratory data analysis

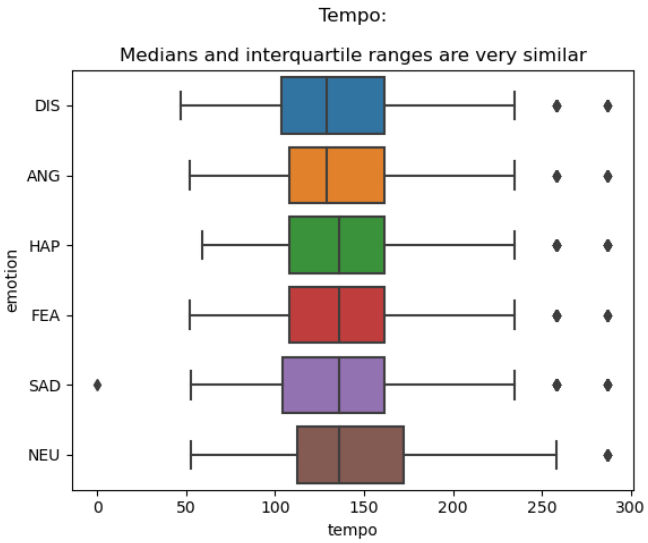
The data was checked for null values and several duplicate recordings were removed. Overall, the dataset was extremely balanced with near equal numbers of observations across each category and at the intersection of categories.

We used an ANOVA test to check whether different emotions had different mean values on key descriptive statistics. Below is a table summarizing the number of six descriptive statistics that had p-values significant at the 0.01 level for each combination of emotions. We can see that some emotions have more significant differences suggesting they are more differentiated.



Notably, fear and happiness may be difficult to distinguish, while anger seems particularly easy to distinguish from happiness and the neutral state.

When comparing the distribution of descriptive statistics across emotions, there was substantial overlap adding to the classification challenge. Below is a comparison of tempo distributions.



Clearly, tempo alone will not allow for effective classification. Generally, all the descriptive statistics had similarly overlapping distributions.

As a final step in EDA, we look at waveforms and spectrograms across the emotions. While they do appear different, it would be extremely difficult to rely on a visual assessment to identify emotion.

Average features and raw audio data

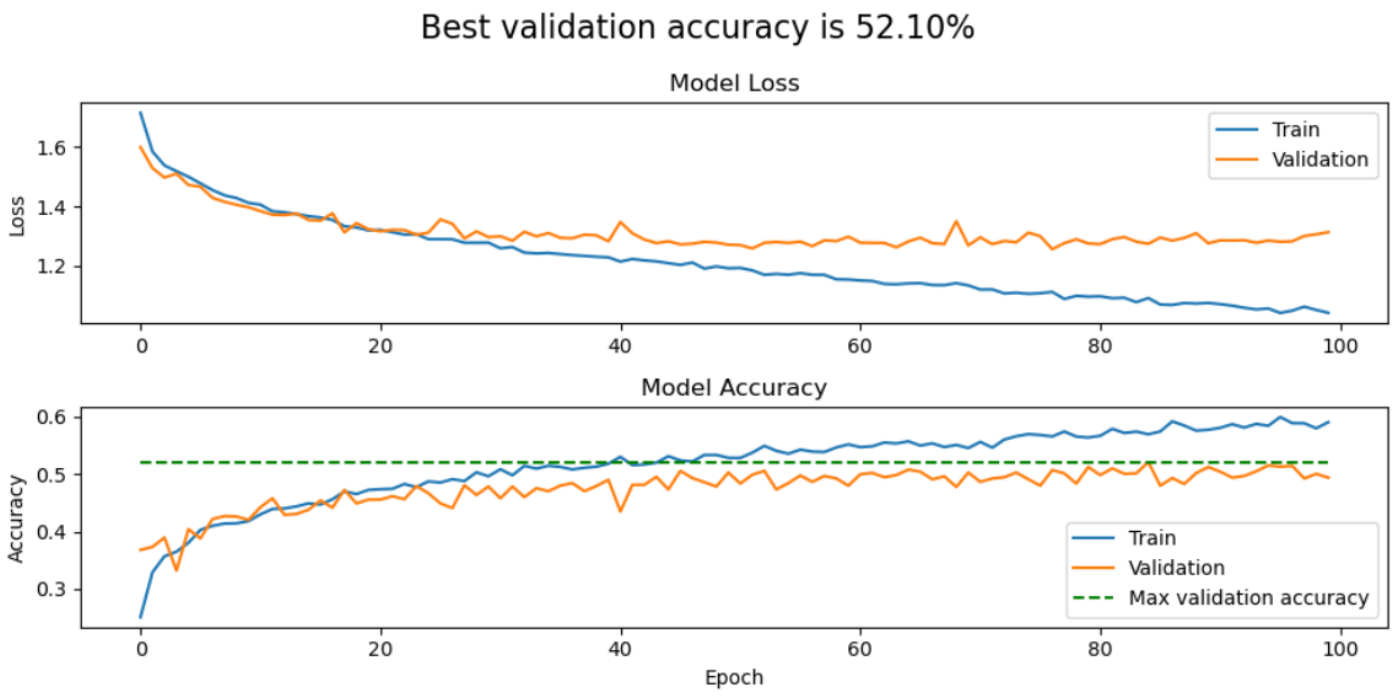
We used two major types of features in our analysis: averages over time of summary features and raw audio data.

For our dense neural network, we used the averages which allow the conversion of audio clips of different lengths into a feature set with a uniform length. Our final feature set included chroma, mels, mfccs, zero-crossing rate and spectral measures.

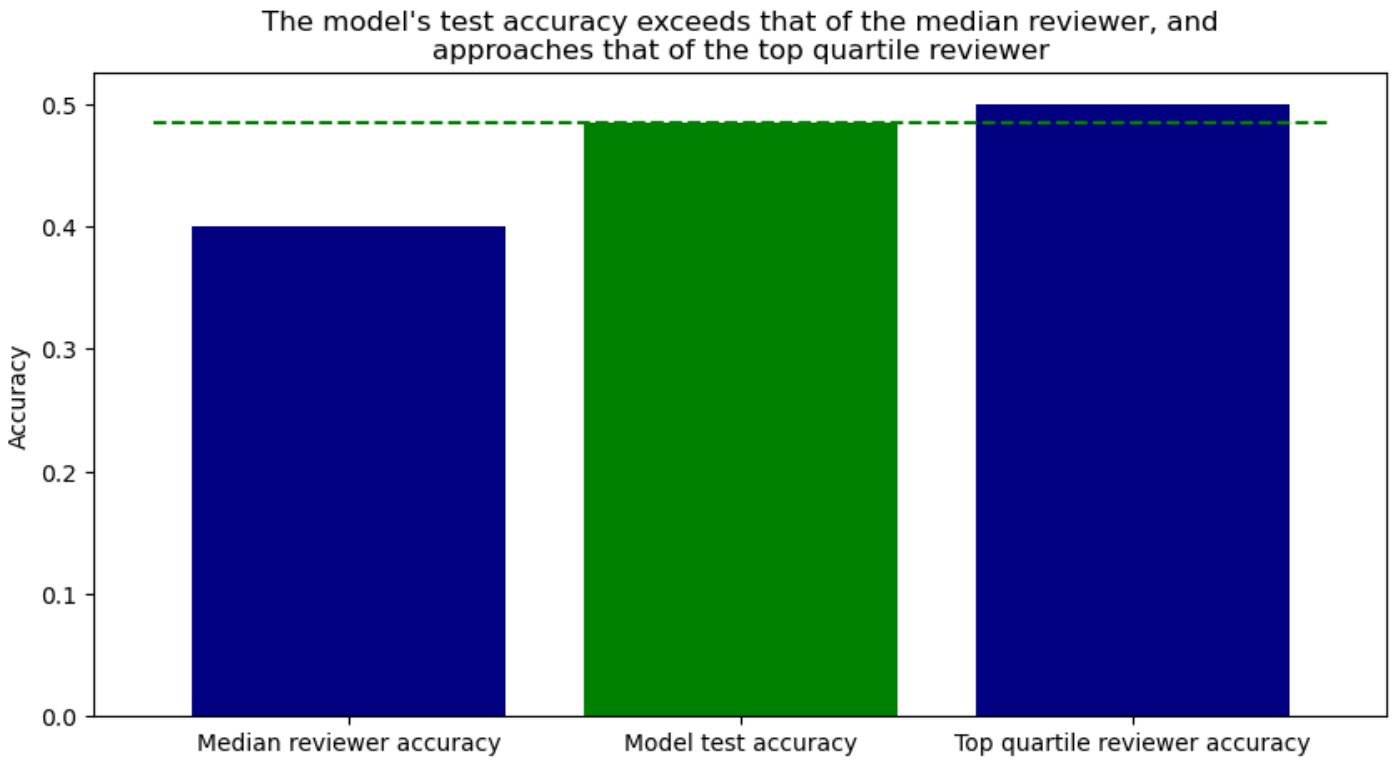
For the convolutional neural network, we used raw audio data with a sample rate of 22,050Hz. We also tested a version downsampled to 8,000hz. To create uniform shapes for our input data, we trimmed off silence from the beginning and end of the audio clips, then padded shorter clips to achieve a uniform length of two seconds. This resulted in observations sampled at 22,050Hz having 44,100 features and observations sampled at 8kHz having 16,000 features.

Dense neural network achieves test accuracy of 48.5%, exceeding the average accuracy of reviewers

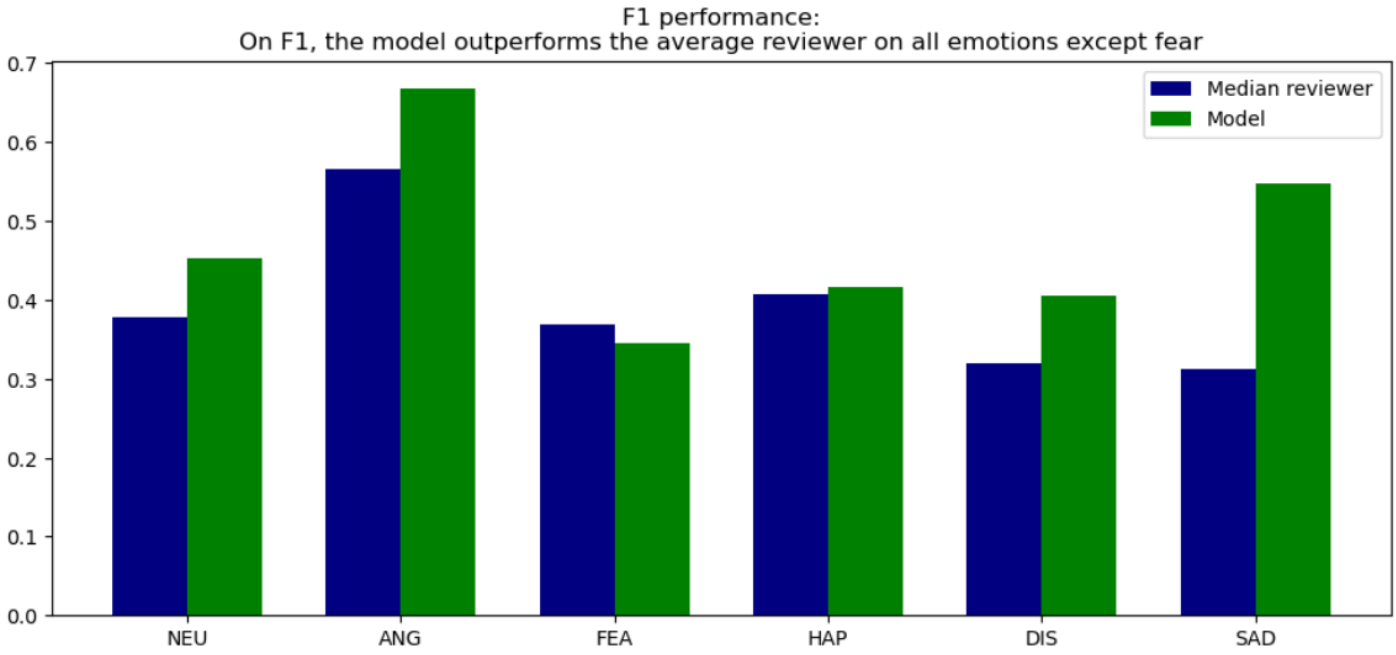
We ran both the dense neural network and one-dimensional convolutional neural network on a variety of parameters. Both the DNN and CNN yield validation accuracies ranging from 45% to over 50%. After running both models on a variety of parameters, we ultimately select the DNN for its higher validation accuracy. When running for 100 epochs, the model shows noticeable overtraining; after 50 epochs, the training accuracy continues to rise while the validation accuracy plateaus. Given this, our final model is trained at just 50 epochs where it is already approaching peak validation accuracy and training accuracy has not moved excessively into the overfitting range. Below is the training history after 100 epochs.



The final model achieves a test accuracy of 48.5%, almost as high as the top quartile of reviewers.



Regarding the specific emotions, the model was much less heavy-handed than the human observations when applying the neutral label. This resulted in improved F1 scores on five of the six classes. The model shows the greatest F1 gains on sadness. Reviewers may have had trouble recognizing this emotion since it has a relatively low intensity compared to representations of other emotions.



Extending the model

While the model significantly exceeds median reviewer performance, test accuracy still does not exceed 50%. In production applications, better accuracy could be achieved by adding information, such as natural language processing to extract word meaning, video imaging to capture body language, and additional context (e.g. customer interaction history). All of these are features that humans typically use when interpreting meaning.

From an audio-only modeling perspective, there may be an opportunity to reduce the number of features used to achieve the current accuracy. Our modeling efforts did not suggest that additional complexity or computing power would lead to better test accuracy; all the model permutations very quickly began overfitting when additional epoch and additional parameters were added.

Our exploration and modeling show that machines can achieve higher accuracy than humans in identifying basic emotions. There are significant opportunities to apply these approaches in production settings to achieve lower-friction interpersonal interactions.

Dataset citation

RAVDESS: The Ryerson Audio-Visual Database of Emotional Speech and Song
- Use citation, academic paper and file download:
  https://zenodo.org/record/1188976

CREMA_D: Crowd-sourced Emotional Multimodal Actors Dataset
- Use citation: Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R.
  CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. IEEE Trans Affect
  Comput. 2014 Oct-Dec;5(4):377-390. doi: 10.1109/TAFFC.2014.2336244.
  PMID: 25653738; PMCID: PMC4313618.
- Academic paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/
- Data download: https://github.com/CheyneyComputerScience/CREMA-D