Machine learning tools can predict emotion in voices in just over a second

March 20 2024

Credit: Pixabay/CC0 Public Domain

Words are important to express ourselves. What we don't say, however, may be even more instrumental in conveying emotions. Humans can often tell how people around them feel through non-verbal cues embedded in our voice.

Now, researchers in Germany have sought to find out if technical tools,

too, can accurately predict emotional undertones in fragments of voice recordings. To do so, they compared three ML models' accuracy to recognize diverse emotions in audio excepts. Their results were published in *Frontiers in Psychology*.

"Here we show that machine learning can be used to recognize emotions from audio clips as short as 1.5 seconds," said the article's first author Hannes Diemerling, a researcher at the Center for Lifespan Psychology at the Max Planck Institute for Human Development. "Our models achieved an accuracy similar to humans when categorizing meaningless sentences with emotional coloring spoken by actors."

Hearing how we feel

The researchers drew nonsensical sentences from two datasets—one Canadian, one German—which allowed them to investigate whether ML models can accurately recognize emotions regardless of language, cultural nuances, and semantic content.

Each clip was shortened to a length of 1.5 seconds, as this is how long humans need to recognize emotion in speech. It is also the shortest possible audio length in which overlapping of emotions can be avoided. The emotions included in the study were joy, anger, sadness, fear, disgust, and neutral.

Based on training data, the researchers generated ML models which worked one of three ways: Deep neural networks (DNNs) are like complex filters that analyze sound components like frequency or pitch—for example when a voice is louder because the speaker is angry—to identify underlying emotions.

Convolutional neural networks (CNNs) scan for patterns in the visual representation of soundtracks, much like identifying emotions from the

rhythm and texture of a voice. The hybrid model (C-DNN) merges both techniques, using both audio and its visual spectrogram to predict emotions. The models then were tested for effectiveness on both datasets.

"We found that DNNs and C-DNNs achieve a better accuracy than only using spectrograms in CNNs," Diemerling said. "Regardless of model, emotion classification was correct with a higher probability than can be achieved through guessing and was comparable to the accuracy of humans."

As good as any human

"We wanted to set our models in a realistic context and used human prediction skills as a benchmark," Diemerling explained. "Had the models outperformed humans, it could mean that there might be patterns that are not recognizable by us." The fact that untrained humans and models performed similarly may mean that both rely on resembling recognition patterns, the researchers said.

The present findings also show that it is possible to develop systems that can instantly interpret emotional cues to provide immediate and intuitive feedback in a wide range of situations. This could lead to scalable, cost-efficient applications in various domains where understanding emotional context is crucial, such as therapy and interpersonal communication technology.

The researchers also pointed to some limitations in their study, for example, that actor-spoken sample sentences may not convey the full spectrum of real, spontaneous emotion. They also said that future work should investigate audio segments that last longer or shorter than 1.5 seconds to find out which duration is optimal for emotion recognition.

More information: Implementing Machine Learning Techniques for Continuous Emotion Prediction from Uniformly Segmented Voice Recordings, *Frontiers in Psychology* (2024). DOI: 10.3389/fpsyg.2024.1300996

Provided by Frontiers