Speech Emotion Recognition Using Deep Learning

Vrisan Dubey [1] Adrian Apsay [2] Luke Taylor [3] Deepika Senthil [4]

[1]vdubey@ucsd.edu [2]adapsay@ucsd.edu [3]lwtaylor@ucsd.edu [4]dsenthil@ucsd.edu

Overview

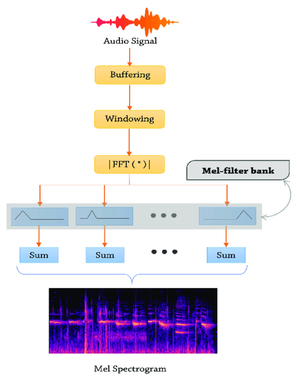
For our project, we aim to classify human speech by emotion through building an audio based machine learning model. We experiment with several techniques in our project, which include using a basic SVM, developing a CNN model from scratch, and deploying a pre-trained transformer SER model.

Key Ideas

Waveform - a graphical/mathematical representation of the shape and characteristics of a signal over time

Fourier Transform - a mathematical operation that decomposes a waveform into a sum of sine waves to reveal its frequency components

Mel spectrogram - a visual representation of the spectrum of frequencies transformed into the Mel scale, which more closely aligns with human auditory perception to ensure accurate representations of perceived frequency.



Dataset

We used the CREMA-D dataset (Crowd-sourced Emotional Multimodal Actors Dataset), which contains recordings from actors across a variety of backgrounds saying short sentences with basic intended emotions. The list of emotions they focused on is as follows:

Emotions:

1. 😐 **Neutral** - most recognized
2. 😁 **Happy** - well identified with one modality
3. 😠 **Anger** - well identified with one modality
4. 🤢 **Disgust** - participants needed audio and visual to identify correctly
5. 😨 **Fear** - participants needed audio and visual to identify correctly
6. 😢 **Sad** - least recognized

Here is some basic information about the dataset:

Clips	Actors	Raters
7,442	91	2,443

Table 1. Data Set Statistics

Audio Only	Visual Only	Audio-Visual
40.9%	58.2%	63.6%

Table 2. Human Emotion Recognition Accuracy Across Modalities

We can see that even human interpreters poorly recognized the actor's intended emotion based on audio alone, as more than half of the time they predicted incorrectly.

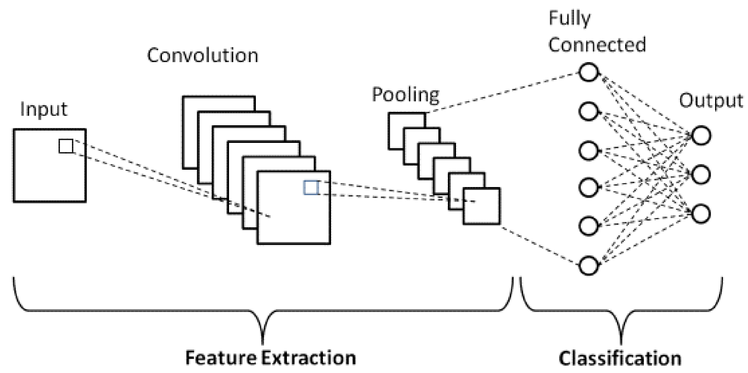
SCAN ME

Methods

Baseline Support Vector Machine

We constructed an SVM for our baseline model. The audio waveforms were processed using FFT to extract the frequency components, on which the SVM classifier was trained with hyperparameter tuning for regularization optimization.

Convolutional Neural Network



We fed each mel-spectrogram for each recording into our CNN. Our CNN consisted of 7 convolutional layers each with ReLU activations, max-pooling, and then dropout to prevent severe overfitting. For our classification task, we added a fully connected, linear layer and then softmax to get our logits. Primary hyperparameters:

number of layers	7
number of epochs	36
initial learning rate	0.0005
learning rate scheduler	Cosine Annealing
optimizer	Adam
dropout rate	0.00165

Table 3. Hyperparameters

Pre-Trained Transformer

We used a Hugging Face pre-trained transformer-based architecture fine-tuned on the CREMA-D dataset, where audio waveforms processed into Mel spectrograms are fed into the transformer layers for feature extraction and emotion classification, utilizing self-attention mechanisms to capture dependencies between audio features across time.

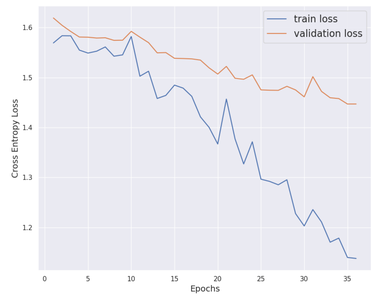
Results

Results across the three models on CREMA-D Dataset:

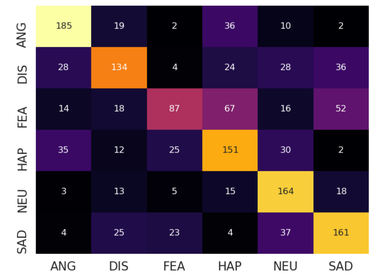
	Accuracy	F1 Score
SVM	0.26	0.24
CNN	0.60	0.59
Pre-trained Transformer	0.73	0.71

Table 4. Model Comparison

CNN Results:



(a) Training and Validation Loss Curves

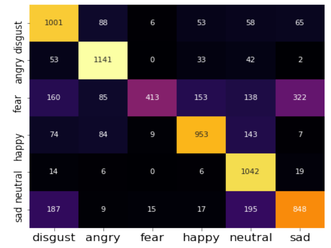


(b) Confusion Matrix of Test Results

Emotion	Precision	Recall	F1-Score	Support
ANG	0.69	0.73	0.71	254
DIS	0.61	0.52	0.56	254
FEA	0.61	0.36	0.45	254
HAP	0.50	0.60	0.55	255
NEU	0.59	0.75	0.66	218
SAD	0.60	0.65	0.62	254

Pre-trained Transformer Results:

Emotion	Precision	Recall	F1-Score	Support
ANG	0.81	0.90	0.85	1271
DIS	0.67	0.79	0.73	1271
FEA	0.93	0.32	0.48	1271
HAP	0.78	0.75	0.77	1271
NEU	0.64	0.96	0.77	1087
SAD	0.67	0.67	0.67	1271



Conclusion

Takeaways and Challenges:

1. Our CNN model was able to outperform the accuracy of human interpreters by almost 20%, while the pre-trained transformer model outperformed by roughly 32%.
2. We outperformed our baseline model by almost 40%, meaning that our feature extraction using frequencies and spectrograms, along with our model development, made a significant difference. The pre-trained transformer model outperformed our baseline model by roughly 47%.
3. Classes like anger, sad, and neutral are predicted well by our model. However, we see a dip in performance in classes like fear. Emotions are pretty much subjective. Someone's *fear* could be someone else's *sadness*. The data were collected by actors trying to express these emotions. The lack of objectivity inherently limits our model.
4. Limited dataset variety poses a challenge for the model to predict emotion on freeform expressions due to the repetitive nature of sentences in our dataset.

https://speech-emotion-classifier.streamlit.app/

Github: https://github.com/vdubey0/ds3-audio