Speech Emotion Classification Multi-Class Prediction Using SVM, CNN, and Transformer Models

Jacob Lamadrid
Cognitive Science, Machine Learning
A16282531

Abstract

The goal of this project is to classify the main emotion of a short segment of spoken speech within a realistic recording setting. The aim is to achieve a predictive accuracy at the level of the average person and to adapt to new recordings of varying sizes for near real-time classification in an implementation of a physical system. According to the original paper of the Crema dataset, "The human recognition of intended emotion for the audio-only, visual-only, and audio-visual data are 40.9%, 58.2% and 63.6% respectively" [1], and therefore these are the central benchmarks of concern. In addition to achieving human capabilities, I am also attempting to match previous benchmarks set by various models [2].

1 Introduction

The task of speech emotion classification is a well-known and well-implemented machine learning task in audio intelligence and as such, several datasets have been developed and used for this task. In this project, the CREMA dataset is used, which consists of crowd-sourced speech recordings of various sentences repeated with different emotions. The dataset was chosen for its lower quality recordings and wider variety, which better generalizes to weaker recordings and helps understand the capabilities of current PyTorch and signal processing libraries in comprehending noisy signals.

Speech emotion classification has diverse utility, such as labeling unlabeled audio data, building larger models for monologue labeling, and certain user interface applications that connect to user messages and intentions. Different models may be more appropriate based on speed, complexity, and generalizability. However, in this project, accuracy is the central metric to determine the effectiveness of models in speech emotion classification.

2 Methods

2.1 Audio Processing and Feature Extraction

The ability to execute convolutions and apply gradients to audio is made possible by our ability to convert audio to a visual form and retrieve information from this form. Conventional approaches to audio classification often utilize the MelSpectrogram format to interpret audio, as it provides a visual representation that is more abstractly connected to our understanding of audibility. However, a step beyond this is the MFCC format, which offers further specificity in audio representation. MFCC (Mel Frequency Cepstral Coefficients) is a representation that emphasizes perceptually relevant aspects of audio across bins. By extracting MFCC features, we can better capture the unique characteristics of audio signals, enabling more accurate classification and analysis.

Much effort has been made in this realm of audio visualization within Python, namely in TorchAudio, which provides great tools for these conversions and provide minimal loss in audio quality relative to

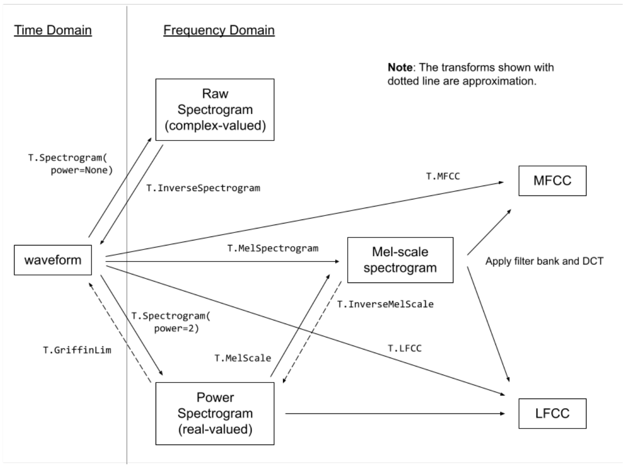

Time Domain
Frequency Domain
Note: The transforms shown with dotted line are approximation.
Raw Spectrogram (complex-valued)
T.Spectrogram(power=None)
T.InverseSpectrogram
T.MFCC
MFCC
waveform
T.MelSpectrogram
Mel-scale spectrogram
Apply filter bank and DCT
T.InverseMelScale
T.Spectrogram(power=2)
T.GriffinLim
T.MelScale
T.LFCC
Power Spectrogram (real-valued)
LFCC

Figure 1: TorchAudio Spectrogram Tools [3]

other packages experimented with. This is where the MFCC function is derived from in this project and more specifically the torchaudio.compliance.kaldi library, which has been specialized for speech.

2.2 Support Vector Machines

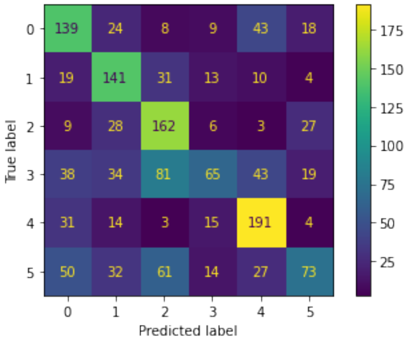
The Support Vector Classifier model from the libsvm library is trained using the summary statistics of the generated MFCCs in the time and frequency domains. This method achieves a classification accuracy of approximately 43%. Although this accuracy is below earlier benchmarks, the SVM model provides a reasonable classification model considering its simplicity and speed.

2.3 Convolutional Neural Networks

2.3.1 ResNet18

A pretrained ResNet18 model from the PyTorch models library is used for audio classification. The ResNet18 architecture is commonly used for this task and achieves an accuracy of 52%.

However, there are significant misclassifications in the disgust and fear classes, accounting for nearly 56% of the errors.



	0	1	2	3	4	5
0	139	24	8	9	43	18
1	19	141	31	13	10	4
2	9	28	162	6	3	27
3	38	34	81	65	43	19
4	31	14	3	15	191	4
5	50	32	61	14	27	73

Figure 2: ResNet18 Confusion Matrix
(0-Happy, 1-Neutral, 2-Sadness, 3-Disgust, 4-Anger, 5-Fear)

2.3.2 VGG16

A pretrained VGG16 model from the PyTorch models library is also used. Although VGG16 is not commonly used for audio classification, it shows promise in this speech emotion recognition problem, achieving an accuracy of 55%.

Again, there are significant misclassifications in the disgust class, accounting for 30% of the errors.

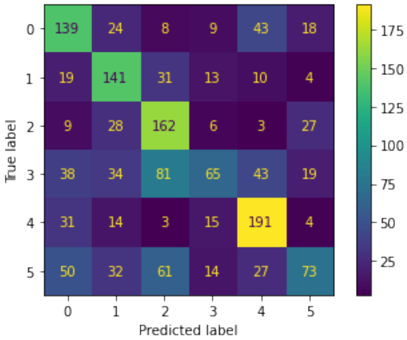


Figure 3: VGG16 Confusion Matrix
(0-Happy, 1-Neutral, 2-Sadness, 3-Disgust, 4-Anger, 5-Fear)

2.4 Visual Transformers

Transformer models are very effective models for visual classification tasks and have been utilized in many contemporary machine learning problems in computer vision. These have also been adapted to audio and in specific, specialized for use on MelSpectrogram representations. However, in this project a pretrained ViT16 from PyTorch's models library is utilized. This model is specifically used in image classification, but it may still be possible to effectively classify audio based on its features in the frequency and time domains. The MFCC representation is used in training.

Despite this models large complexity in comparison to the others tested, the model is only able to achieve an accuracy of 42%. This is likely due to the choice of audio representation and its forceful resizing of the image. Due to this, it may be better suited for representations outside of MFCC and MelSpectrogram as both have been trialed and achieved low accuracies. The dimensionality of this Transformer model most likely must be altered in order to achieve a reasonable accuracy.

3 Experiments

3.1 Audio Representations

Various audio representations were trialed which includes the MelSpectrogram, Spectrogram, MFCC, and Waveform images of the audio. Among these, the MFCC representation performed best across models due to its inherent feature extraction and specificity across frequencies. The MelSpectrogram is typically the preferred representation, but, likely due to zero-padding and reshaping, its performance remained low and the MFCC provided accuracies consistent with other research in this field made publicly available.

3.2 Optimizers and Parameters

Used by all of these models is the Adam optimizer at a learning rate of 1e-5. This was the result of testing each model on a smaller subset of the data and a learning rate of 1e-3, 1e-4, and 1e-5 with the Adam, Adagrad, and SGD optimizers. Among all the models requiring an optimizer, the Adam optimizer provided the best accuracy and thus was chosen. The learning rate worked best at such a small number due to the properties of the Adam optimizer and its gradient computation, but also due to the models being pretrained, thus needing less adjustments over less iterations. As such, each model was optimized over only 6 epochs.

4 Conclusion

From the results of experimentation and model development, we see a clear pattern in the failure to accurately classify complex emotions found within speech, more specifically disgust and fear as seen in our dataset. However, we see high accuracies across other, more basic emotions.

In terms of models, the VGG16 performed most effectively having the highest accuracy among those tested and being trained within a reasonable timeframe. This along with the ResNet18 performed the best showing the effectiveness of Convolutional Neural Networks in audio classification tasks. However, much work has been made in Audio Transformers and although the the low performance of the PyTorch Visual Transformer model, Audio Transformers have proven to be effective classification models over CNNs in other tasks outside of speech emotion classifcation.

The overall accuracy of the models utilized was fairly low, but in regards to the lower quality of audio recording and massive variability due to crowd sourced data, this accuracy is to be expected from baseline SVM, CNN, and Transformer models as many models which acheive greater accuracies include very complex aspects in addition to the base model. Therefore, while there is a lack of complexity in models, the lower accuracy is also a sacrifice resultant from the training data in favor of a more generalizable model which may be more successfully implemented in future projects in which realistic audio recordings are of concern.

5 Supplementary Material

Code, and demo can be found here: github.com/j-lamadrid/speech-emotion-classification

References

[1] CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. Nih.gov. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313618/

[2] 1505.07376 Papers with code - crema-D benchmark (speech emotion recognition)
The latest in Machine Learning. (n.d.) https://paperswithcode.com/sota/speech-emotion-recognition-on-crema-d

[3] torchaudio.transforms - torchaudio 2.0.1 documentation Torchaudio.transforms¶, https://pytorch.org/audio/stable/transforms.html, torchaudio.transforms - Torchaudio 2.0.1 documentation