Auditory Unveil: Decoding Emotions in Speech

George Contreras Oscar Lay Thejaswin Kumaran Harper Wood Leila Igwegbe
Saanvi Bala Dr. Yapeng Tian

Introduction

In this project, we **transform audio signals into visual representations**, reading them as images through the utilization of **Mel-frequency cepstral coefficients (MFCCs)**. The **Convolutional Neural Network (CNN)**, trained on these visualized audio patterns, learns to discern emotions by recognizing features associated with various emotional states. The integrated use of audio and image processing significantly enhances the accuracy of the overall project. We managed to find a distinctive approach to speech emotion detection by employing image classification techniques typically applied to visual data.

Data Collection

We trained our models using a combination of 4 popular datasets (RAVDESS, CREMA-D, TESS and SAVEE) containing English phrases stated in different tones. All the datasets had collections of voice recordings, saved as .wav files, that we compiled according to the tone of the speech (angry, disgust, fear, happy, neutral or sad). Using the Fourier transform, all the recordings are then converted into spectrograms, allowing us to visualize audio files on a graph. The graph shows the amplitude of the recording at each frequency as time goes on. The specific kind of spectrogram we use are MEL diagrams where Hz is represented on a logarithmic scale.

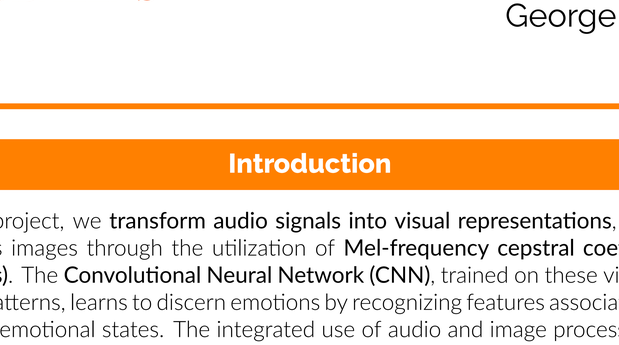

Figure 1. MEL Diagram made of an angry tone

We have about 11,741 total recordings, many with different scripts, tones, accents, and voice actors to ensure variety. Some recordings however have the same actor, tone, and script, but different pronunciations of the same words, which can help the model avoid overfitting. This ensures that the model trains towards tone-detection, as opposed to the mannerisms of the actor.

Model

- **Supervised Learning**: When a model is trained on a labeled data set where for each input value, the output value is known. The model's weights are adjusted until the model is fitted, which means that it produces the desired outputs.
- **Convolutional Neural Networks (CNNs)**: CNNs are made up of convolutional layers and pooling layers. In each convolutional layer, a filter passes over each section of an image to check for a particular feature. The pooling layers reduce the sizes of the feature maps produced by the convolutional layers. This architecture is especially effective for image classification.
- **Our Model**: Using TensorFlow and Keras, we developed a 2D CNN for image classification of spectrograms generated from the audio data. The neural network identifies features in the spectrograms which correspond to different emotions. Our model utilized supervised learning; for each audio file in the training data set, the emotion was known, and the model was adjusted to output the correct emotions. We also refined the model to prevent overfitting.

Analysis

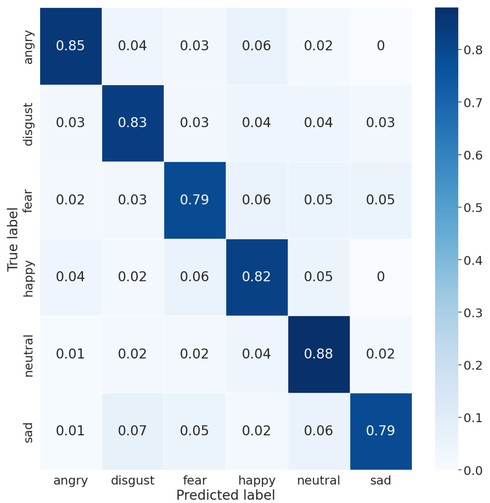

Figure 2. Confusion Matrix

Results

The model, upon its initial evaluation, demonstrated a modest accuracy of 51.74% in emotion classification. Following the implementation of lasso and ridge regularization, dropout regularization, data normalization, and more 2D convolution layers, the accuracy skyrocketed to an impressive 82.68%. This substantial increase showcases the effectiveness of the proposed enhancements in accurately discerning human emotions from audio files, marking a notable breakthrough in the field.

Conclusion

Our project proves that it is possible for a machine to detect emotion from something as simple as our tone. This could allow us to interact with machines in increasingly personal ways. We only interact with computers through text or speech-to-text, which only carries some context. It's hard enough to talk to other people in written format, requiring emojis and tone indicators to get a full context of another's speech. With tone-detection, machines could respond more appropriately. For example, when a user seems consistantly sad/depressed, an assistant could try to console them. In a more healthcare or law-enforcment related context, emotion detection could provide a deeper understanding of patients' emotional states during remote consultations or calls, aiding first-responders in gauging urgency or de-escelation. This model serves as a foundation for additional research, and with a larger and more inclusive dataset, this technology has the potential to change how our machines can interpret us.

References

Alam, Mahbub. (2020, December 13). *Data normalization in machine learning*. Medium. https://towardsdatascience.com/data-normalization-in-machine-learning-395fdec69d02

Bhattacharyya, S. (2023, September 18). *Ridge and lasso regression: L1 and L2 regularization*. Medium. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

Keras Team. (n.d.). *Keras documentation: Image data loading*. https://keras.io/api/data_loading/image/

Load and preprocess images. (n.d). TensorFlow. https://www.tensorflow.org/tutorials/load_data/images

What are convolutional neural networks?. (n.d.). IBM. https://www.ibm.com/topics/convolutional-neural-networks