Speech Emotion Recognition

Habib Unluoglu[1], Sania Eskandari[2], Brooke McWilliams[3], James Birch[4]

[1]Student ID: 12480310; email: hun222@uky.edu

[2]Student ID: 12478140; email: ses235@uky.edu

[3]Student ID: 12543131; email: brmc230@uky.edu

[4]Student ID: 12293085; email: jabi236@uky.edu

Abstract: Emotions are crucial in human interaction, and accurately identifying them from speech signals poses a challenging yet essential task. In this report, we present a comprehensive exploration into speech emotion recognition (SER) utilizing machine learning methods. Our project report concentrates on feature extraction, selection, and classification methodologies to differentiate the various emotional states conveyed through speech. We explore various acoustic features and extract relevant information from the speech signal. Feature selection techniques are employed to identify the most discriminative features and optimize the model performance, enhancing the efficiency and effectiveness of subsequent classification models. We employ Convolutional Neural Networks (CNN) machine learning algorithm in order to classify emotions. The model is trained and validated on "Crema" dataset containing labeled emotional speech samples. Evaluation metrics such as accuracy are applied to assess the model performance. Our results demonstrate promising accuracies in recognizing a range of emotions, including sad, anger, disgust, fear, happy and neutral states. Additionally, we discuss the challenges encountered, such as dataset pre-processing, data variability and model generalization. This report provides valuable insights into the field of speech emotion recognition, presenting illumination on successful methodologies and their potential applications across various domains.

GitHub Link: brmc230/CS460G-Project: Speech Emotion Recognition (github.com)

1. INTRODUCTION

Numerous forms of communication exist, but the speech signal stands out as one of the fastest and most instinctive means of human interaction. Consequently, speech can also serve as a rapid and effective mode of communication between humans and machines (Ayadi et al. 2011). Expressing emotions and attitude through language is also important. Specifically, detecting emotional cues in speech signals and determining the underlying emotions within spoken expressions is a significant and valuable endeavor

(Swain et al. 2018). Therefore, due to the digitally growing era and the importance of emotions within spoken expressions, speech emotion recognition holds substantial importance in a wide array of applications, including Human-Computer Interaction (SCI), lie detection, steering assistance in automotive systems, intelligent tutoring, audio mining, security, telecommunications, and human-machine interaction in various settings like homes, hospitals and shops (Koduru et al. 2020).

We, as a group, have chosen to embark on our group project in the domain of "speech emotion recognition" for specific reasons. One of the primary motivations behind our choice is the growing importance and relevance of this field in various applications and industries. Additionally, we believe that our exploration of speech emotion recognition will not just deepen our comprehension of the technology but will also grant us valuable perspectives into the ever-changing realm of human-computer interaction, affective computing, and the prospects for enhancing user experiences across diverse scenarios. By embarking on this project, we aim to contribute to the body of knowledge in this area and develop practical applications that can make a positive impact in fields. A challenge for this project will be assuring that the data is not skewed and stays uniform after extracting the features from each file. Male voices could give higher feature values while female voices could give lower feature values but result in the same class label (emotion).

2. DATASET

In this report, we opted to use the "Crema" dataset specifically for our project report centered on speech emotion recognition, acknowledging its relevance in analyzing emotional expressions communicated through speech. The dataset encompasses various emotional states, each corresponding to specific emotion labels:

SAD: corresponds to the emotional state of sadness

ANG: corresponds to the expression of anger

DIS: corresponds to the emotion of disgust

FEA: corresponds to the state of fear

HAP: corresponds to happiness or joy

NEU: corresponds to a neutral emotion state

Each emotion category in the Crema dataset corresponds to audio recordings featuring speech samples expressing specific emotions and includes a total of 7442 audio files. This diversity enables the

development and evaluation of models focused on precisely identifying and distinguishing different emotional states communicated through speech signals. Utilizing the Crema dataset, emotion recognition systems play a substantial role in various applications, such as affective computing, human-computer interaction, and assessing mental health.

3. METHODOLOGY

In this report, we utilize the Convolutional Neural Networks (CNN) machine learning technique to categorize emotions. Although CNNs were originally designed for image analysis, CNNs have shown versatility in handling sequential data like audio spectrograms, making them a fitting choice for our SER task. We adapted the CNN model to process speech spectrogram representations, enabling it to autonomously grasp essential features at various abstraction levels. Through multiple layers of convolution and pooling, the network captured intricate patterns and temporal relationships within the audio data, enhancing its capacity to discern subtle emotional expressions.

Our methodology involved a process of data preprocessing, transforming raw audio samples into spectrograms, which served as input the CNN. Subsequently, the model underwent iterative training and fine-tuning stages. During training, CNN learned to extract features from the spectrograms while iteratively adjusting its internal parameters to minimize classification errors.

Hyperparameters were fine-tuned, including kernel sizes, stride lengths, and network depth, ensuring the model's efficacy in classifying emotions across the spectrum. Validation and testing on the "Crema" dataset enabled us to assess CNN's ability to accurately identify and categorize emotional states conveyed through speech.

The robustness and adaptability of the CNN architecture, combined with meticulous training on labeled emotional speech samples, positioned our approach to effectively capture and classify various emotional expressions embedded within speech signals.

4. EXPREMENTAL SETUP

In order to perform comprehensive experimentation for speech emotion recognition using the Crema dataset, a Convolutional Neural Network (CNN) architecture was established as the foundational model.

Model Architecture:

- Sequential model construction

- Conv1d layers: 256 (input shape layer) - 256 - 128 - 64 with 5 kernels and padding
- ReLu activation function
- Max pooling with a pool size of 3 after each activation to account for more features
- Dropout of a quarter of samples after the 3rd Conv1d layer
- Flatten layer
- Dense Layer of 32 units with ReLu activation
- Dropout layer of a quarter of samples
- Dense Layer for the number of labels with softmax activation

Optimization Parameters:

- Optimizer: Adam
- Loss function: sparse_categorical_crossentropy
- Learning rate: Initially set to 0.001, later changed to 0.0001 for improved accuracy

Model Training and Alterations:

- Initial baseline accuracy was 16%
- Experimented with different learning rates, initially changing from the default Adam rate to 0.001
- Started testing with MFCC features, then transitioned to RMS features due to prolonged testing and stagnant accuracy; baseline accuracy with MFCC was 30%
- Added Dropouts and experimented with various dropout factors
- Explored different batch sizes; 16 worked best for the number of features
- Adjusted pool sizes to accommodate more features after further normalization of data
- Altered batch size to 64 to account for more features and improve model speed
- Increased epochs from the initial 33, planning to extend to 50 or 100
- Removed Batch normalization as feature extraction normalized data
- Added more Max Pooling layers to focus on prominent features
- Simplified the model by removing Conv1D layer
- Adjusted the first dense layer to 32 units
- Applied padding to pooling layers

Evaluation Metrics:

- Initial baseline accuracy was 16%, significantly improved to 48.62%.

Software and Tools:

- **Programming Languages:** Python
- **Libraries:** TensorFlow, Librosa, Pandas, NumPy, Matplotlib

5. RESULTS

In our group project on speech emotion recognition, employing the CNN method yielded notable improvements on the Crema dataset to classify emotions. The initial baseline accuracy stood at 16%. After updating the model to accommodate new normalized data through a series of iterative adjustments

and experiments, we observed a significant leap in accuracy, currently achieving around 48%. The progression of changes and their impact on the model's accuracy is summarized below:

1. **Hyperparameter Tuning:** The learning rate was modified from Adam's default to 0.0001, resulting in slight performance improvements.
2. **Architecture Modifications:** Various architectural changes were made, including the removal of final Batch Normalization, addition of Dropouts with different factors, adjustments in batch sizes, pool sizes, and padding strategies.
3. **Feature Engineering:** Initially testing with MFCC features resulted in prolonged testing times without significant accuracy changes. Switching to Root Mean Square Error (RMS) features helped stabilize accuracy at 30% initially.
4. **Normalization and Complexity:** Data was further normalized, allowing for adjustments in pool sizes and the removal of Batch Normalization. Complexity reduction strategies, such as removing Conv1D layers and altering dense layer units to 32, were implemented.
5. **Training Optimization:** Additional Max Pooling layers were added to identify prominent features. The number of epochs was increased to 100.
6. **Current Accuracy:** The final accuracy achieved stands at 48.62%, showcasing a significant improvement from the initial baseline of 16%.
7. **Toolset and Framework:** The experiments were conducted using Python, leveraging TensorFlow, Librosa, Pandas, NumPy, and Matplotlib libraries for model development, analysis, and visualization.

Throughout our experimentation, we encountered **challenges** with overfitting, which necessitated adjustments to the model architecture. By addressing these issues and fine-tuning the model, we achieved a more reliable performance, showcasing the effectiveness of our adapted approach in handling and improving upon the initial limitations. As a result, the iterative adjustments to the CNN architecture, feature engineering, and training optimization strategies contributed to a notable enhancement in model accuracy from the baseline on the Crema dataset. Further experimentation with increased epochs and potential architectural refinements may continue to improve the model's performance.

6. DISCUSSION

In this section, we delve into the challenges encountered during the task of classifying speech emotions.

The undertaken experimentation and refinements in the Convolutional Neural Network (CNN) model for audio classification on the Crema dataset have showcased the significance of iterative model adjustments in enhancing classification accuracy. The incremental modifications in hyperparameters, architecture, and feature engineering strategies played a pivotal role in elevating the model's performance from an initial baseline accuracy of 16% to a commendable 48.62%. The exploration of various learning rates, dropout factors, batch sizes, and normalization techniques underscored the importance of fine-tuning these parameters for optimizing model performance. The observed incremental improvement from a baseline accuracy of 16% to 48.62% suggests *potential challenges* in feature extraction techniques, specifically in determining optimal frame lengths and hop lengths to extract the most informative features for model training. Additionally, the balance between model complexity and performance raises questions regarding further exploration into feature representation and potential simplification or refinement of the model architecture to enhance classification accuracy.

Moreover, the strategic alterations in the model's architecture, such as the addition and removal of layers, padding strategies, and the introduction of more pooling layers, demonstrated the intricate balance between model complexity and performance. The shift from MFCC to RMS features, though initially time-consuming, proved pivotal in stabilizing accuracy at a higher level. Notably, the removal of Batch Normalization and the reliance on feature extraction for data normalization were pivotal in achieving improved accuracy while simplifying the model architecture.

However, the project has inherent limitations. The continual increase in epochs and potential architectural modifications may lead to the risk of overfitting or computational complexities. Additionally, the constraint on pool sizes due to the number of features poses limitations on the model's capacity to discern intricate patterns.

The toolset comprising Python, TensorFlow, and associated libraries served as robust frameworks for model development and analysis. Despite the notable progress in accuracy, there's room for further exploration and refinement, particularly in the feature extraction process, model architecture, and potentially exploring ensemble methods or transfer learning techniques to elevate accuracy beyond the achieved threshold.

To summarize, this project's findings emphasize the importance of meticulous parameter tuning, architectural adjustments, and feature engineering in optimizing CNN models for audio classification tasks. The attained accuracy of 48.62% serves as a promising milestone, yet further research and exploration into advanced methodologies remain imperative for pushing the performance boundaries in audio classification tasks.

7. CONCLUSION

In conclusion, we summarize our achievements, providing an overview of our progress, key findings, and considerations for future directions.

This report outlines our efforts to enhance a specialized computer model (CNN) for recognizing emotions in speech using the Crema dataset. Through various adjustments in how the model learns, its structure, and the way it interprets data features, we successfully increased its accuracy from an initial 16% to a solid 48.62%. However, we encountered challenges in determining the most effective features from the data, significantly impacting the model's performance.

Despite encountering challenges such as overfitting, our endeavor showcased the importance of data preprocessing and iterative model adjustments in enhancing performance. Balancing the complexity of the model with its accuracy was a noteworthy challenge. This led us to consider how we represent features and whether simplifying the model's design could maintain its accuracy. Although reaching an accuracy of 48.62% was a significant achievement, the complexity inherent in decoding emotional nuances from speech signals remains a persistent challenge, indicating the need for further refinement and exploration in algorithmic approaches and model architectures.

Our findings underscore the potential for advancements in speech emotion recognition and highlight the ongoing necessity for more sophisticated methodologies to achieve higher accuracy rates in discerning and classifying emotional states within speech. This project contributes to the collective understanding of the intricate nature of speech emotion recognition and sets the stage for future research aimed at refining models and algorithms in this evolving domain.

8. REFERENCES

Ayadi, M.M., Kamel, M.S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit., 44, 572-587.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. Chapter 10.

Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. International Journal of Speech Technology, 23(1), 45-55.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. International Journal of Speech Technology, 21, 93-120.