Alina KHANOVA MADS

Emotion Classification from Speech

1. Introduction

Emotion recognition from speech is a field that focuses on automatically identifying human emotions based on speech signals. Humans naturally rely on various cues, such as facial expressions, verbal expressions, and body language, to understand emotions. However, developing an automatic system that can accurately recognize emotions from speech poses a challenge. With the vast amount of multimedia information available and the advancements in computational methodologies related to machine learning, the scientific community has dedicated significant effort to advance emotion recognition from speech signals [1, 2].

The topic of speech recognition of emotions is of great importance because of its potential applications in various fields. Accurate emotion recognition can improve human-computer interaction, improve customer service, provide a personalized learning experience, and contribute to the development of intelligent virtual assistants. Emotion recognition systems may also be useful in healthcare, where they can aid in the early detection of mental health conditions and facilitate more effective therapy sessions. In addition, understanding the emotions expressed in speech can provide valuable information about human behavior and promote better communication and understanding between people.

The task considered in this study is to develop an efficient and reliable emotion recognition system based on speech signals. The challenge is to accurately capture and interpret emotional cues from speech, which often contains unstructured and informal information. The goal is to develop a system that can accurately recognize emotions from speech in real time, providing practical applications in various fields.

The hypothesis of this study is that with the help of recurrent neural networks, in particular LSTM models, it is possible to improve the accuracy and reliability of speech emotion recognition. The integration of contextual information and the ability of LSTMs to capture temporal dynamics in speech signals can enhance the system's ability to recognize and classify emotions. The hypothesis suggests that LSTMs can efficiently learn patterns and dependencies in speech data, resulting in performance improvements over traditional approaches.

2. Background

The paper [3] conducted an experimental study on recognizing emotions from human speech, focusing on emotions such as neutral, anger, joy, and sadness. The main finding of the study was that considering data from an individual subject rather than a group of people leads to better accuracy in emotion classification. The limitations of the work include a limited number of emotions considered and a relatively small dataset. Our project will address these limitations by expanding the range of emotions felt and collecting a more extensive and diverse dataset for emotion recognition. Additionally, we will explore advanced techniques such as deep learning models to improve accuracy and performance.

The work in [4] aims to address the limited body of work on emotion detection from speech and investigates the relationship between gender and the emotional content of speech. The main finding of the paper is that certain features extracted from speech, such as pitch, Mel Frequency Cepstral Coefficients (MFCCs), and formants, can carry emotional information and contribute to accurate emotion recognition. The limitations of the work include the debate on which features influence emotion recognition in speech, the uncertainty regarding the best algorithm for classification, and the grouping of emotions together.

The paper [5] aimed to recognize speech emotion by learning deep emotion features using two convolutional neural networks and long short-term memory (CNN LSTM) networks, one for raw audio and one for log-Mel spectrogram data. The main finding of the paper was that the designed CNN LSTM networks achieved excellent performance in recognizing speech emotion, outperforming traditional approaches on benchmark databases. The limitations of the work include the lack of explanation for how the features are learned by the networks and the need for further improvement in several aspects of the designed networks.

The paper by Pandey [6] focuses on deep learning algorithms and their application in analyzing speech data to identify emotional states. The authors explore the use of algorithms such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for evaluating the capacity of emotion detection from standard speech representations, including the magnitude spectrogram, Mel spectrogram, and Mel-Frequency Cepstral Coefficients (MFCCs).

The researchers conducted their experiments on two publicly available datasets, namely the EMO-DB and IEMOCAP datasets. They applied various models and feature combinations to these datasets and presented the results of their studies. The paper provides insights into the rationale behind selecting specific models and feature combinations and aims to determine the optimal approaches for speech emotion recognition.

Overall, Pandey contributes to the field of speech emotion recognition by exploring deep learning techniques, evaluating different models, and examining various feature representations. Their findings shed light on the effectiveness of CNN and LSTM networks and highlight the importance of selecting appropriate features for accurate emotion detection in speech data.

3. Materials and methods

3. 1. Dataset

The dataset used for training the emotion classifier is the TESS (Toronto Emotional Speech Set) dataset. This dataset consists of audio recordings of two actresses, aged 26 and 64 years, speaking 200 target words in each of seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The dataset contains a total of 2,800 audio files in WAV format.

The data was collected by recording the actresses speaking the target words with different emotional expressions. Each emotion category has 400 audio files, resulting in a balanced dataset. The dataset is organized into folders based on the actress and emotion category. The dataset does not provide information on the gender distribution or other demographics beyond the age and emotion categories.

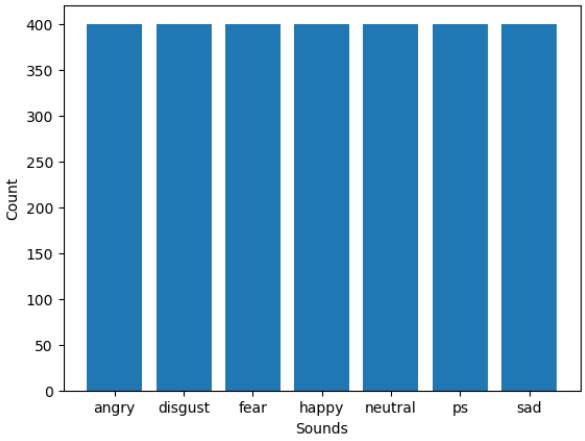

Figure 1: The distribution of the type of emotions by the number of records

Audio quality in the dataset is good, making it easy for humans to judge the emotions. High-quality audio is indeed beneficial for accurately perceiving and categorizing emotions. However, it's important to consider the generalizability of the model trained on such a dataset when it encounters audio with different qualities or characteristics.

MFCC is one method for feature extraction that is be used to analyze speech by extracting critical data and features from subsets of the speech data. The MFCCs of an audio

signal are a minimal number of features (often 10–20) that succinctly reflect the pattern of a spectral envelope. MFCC features are computed using linearly spaced frequency filters at low frequencies and logarithmically spaced frequency filters at high frequencies. As shown in Figure 2,3,4,5 the MFCC matrix for every audio sample is calculated.

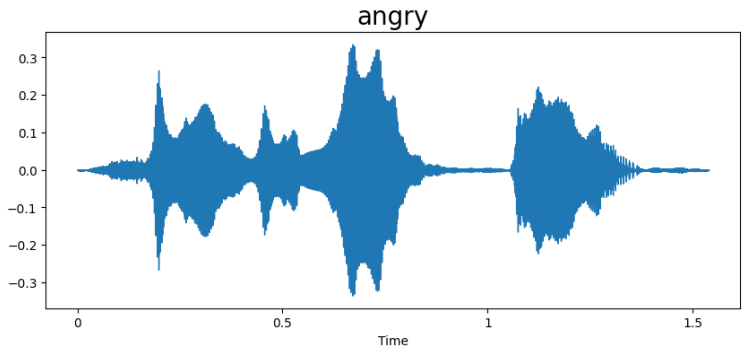

Figure 2: RAVDESS Angry Waveform

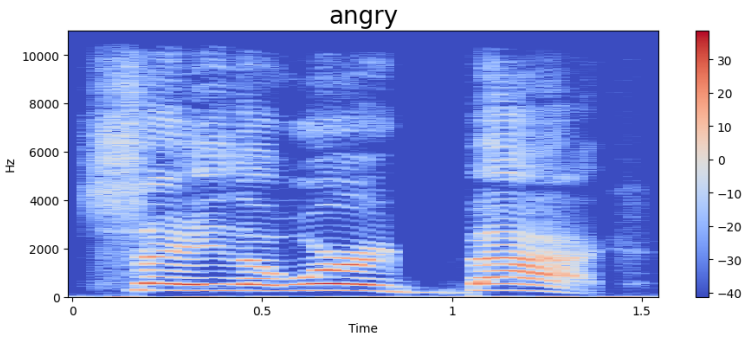

Figure 3: RAVDESS Angry MFCC

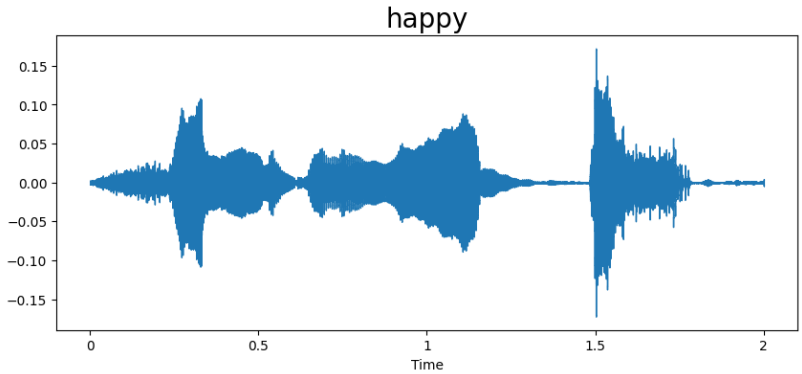

Figure 4: RAVDESS Happy Waveform

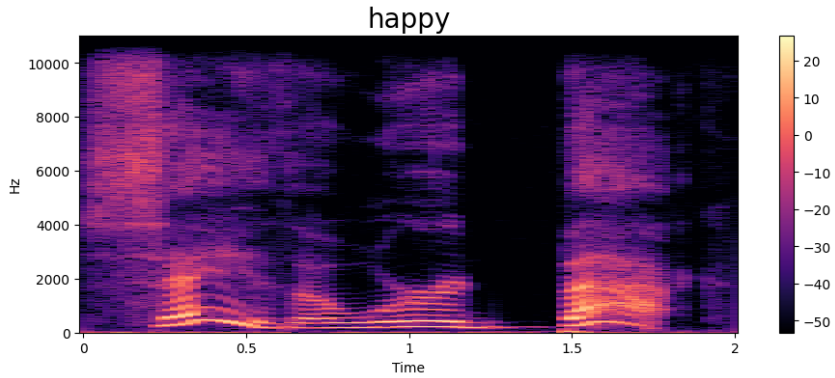

Figure 5: RAVDESS Happy MFCC

3. 2. Method

The model used for training the emotion classifier is a LSTM (Long Short-Term Memory) model, which is a type of recurrent neural network (RNN) commonly used for sequence data. The input to the model is a sequence of MFCC (Mel-Frequency Cepstral Coefficients) features extracted from the audio files. The MFCC features are extracted using the Librosa library.

The model starts with an LSTM layer with 256 units and return_sequences=False (meaning it returns the last output only). The input shape is specified as (40, 1). After the LSTM layer, a dropout layer is added with a dropout rate of 0.2. Dropout is a regularization technique used to prevent overfitting in neural networks. Then, two dense layers with 128 and 64 units, respectively, are added. Each dense layer is followed by a dropout layer. Finally, a dense layer with 7 units (corresponding to the number of classes) and softmax activation is added.

The model is trained using the Adam optimizer with a categorical cross-entropy loss function.

The validation_split argument splits the training data into a validation set (20% in this case) for monitoring the model's performance during training. The model is trained for 50 epochs with a batch size of 64.

4. Results and Discussion

To report the results achieved using the LSTM model, we will evaluate the model's performance on both the training and validation datasets. We will use relevant evaluation metrics, such as accuracy, and provide supporting material such as plots and figures. First, let's analyze the accuracy of the model during training and validation over the epochs:

Plot 6 shows the training and validation accuracy values over the epochs. This plot helps us understand how the model's accuracy improves or converges over time.

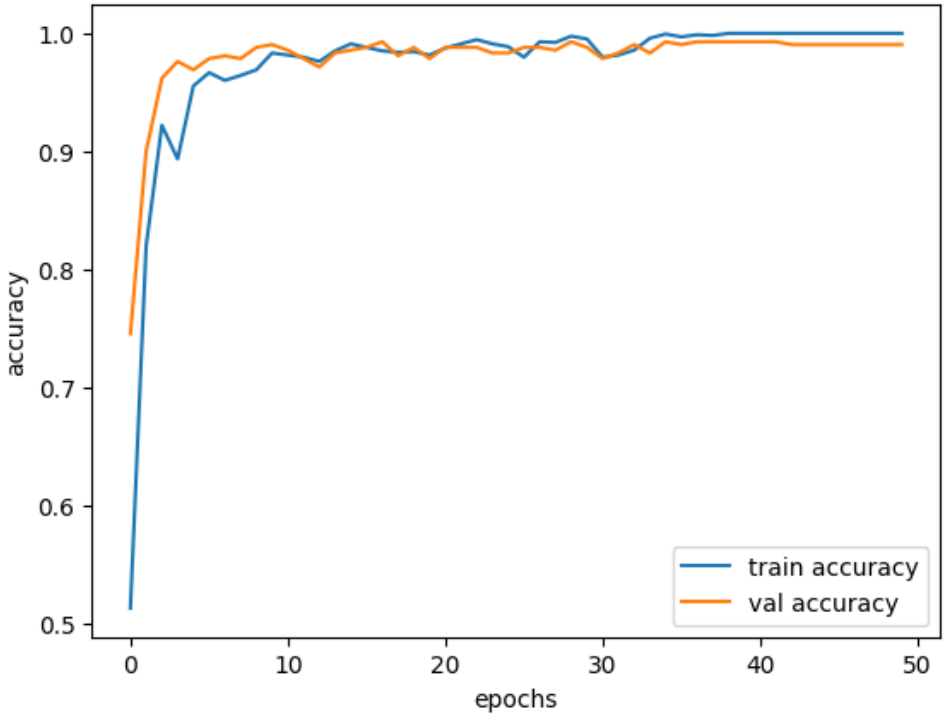

Figure 6: Training and validation accuracy values over the epochs

Next, let's evaluate the model's performance using other evaluation metrics such as loss and plot the results.

Plot 7 shows the training and validation loss values over the epochs. It helps us understand how the loss decreases or converges during training.

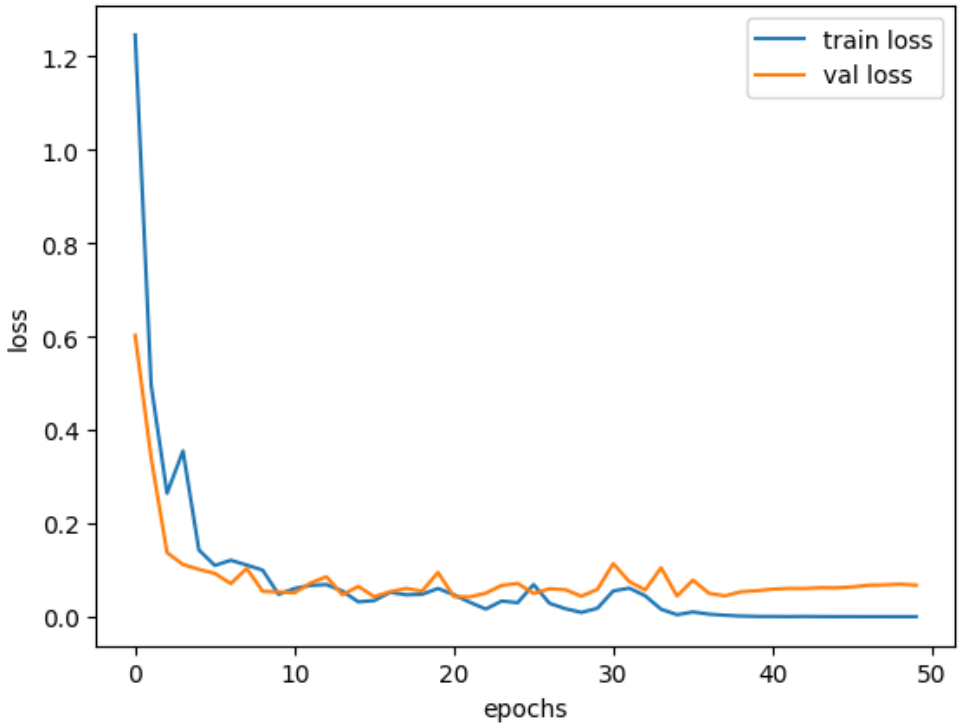

Figure 7: Training and validation loss values over the epochs

From the loss and accuracy plots, it can be observed that the model quickly converges and achieves high accuracy on both the training and validation sets. The loss decreases rapidly in the initial epochs and then stabilizes at a low value. The accuracy increases steadily and reaches a high value, indicating that the model successfully learns to classify the data.

The model achieved an accuracy of 99.88% on the training set and 99.29% on the validation set. The high accuracy indicates that the LSTM model is able to effectively learn the patterns in the input data and make accurate predictions. The model shows good generalization performance, as the validation accuracy is close to the training accuracy.

However, it's important to note that these results are specific to the given dataset and may not generalize to other datasets or real-world scenarios. Further evaluation and testing on additional datasets are necessary to assess the model's performance in different contexts.

Although the LSTM model gave good results, you can try other models, for example, add more LSTM layers or use bidirectional LSTMs. Different architectures can more effectively capture temporal dependencies in the data and potentially improve model performance.

To apply the emotion recognition model in real life, we can implement in real time: an adaptation of the LSTM model for real-time activity recognition applications where it is necessary to make real-time predictions using streaming sensor data.Начало формы

5. Conclusion

In this article, we have considered the task of classifying emotions from speech signals using the long-term short-term memory model (LSTM).

The main hypothesis of work was that LSTMs can improve the accuracy and reliability of speech emotion recognition by capturing temporal dynamics and integrating contextual information. We explored the use of the TESS dataset, which consists of audio recordings of seven emotions uttered by two actresses. We extracted Mel-Frequency Cepstral Coefficients (MFCC) as features from speech data and trained an LSTM model on this dataset.

The results demonstrated the effectiveness of the LSTM model in recognizing emotions from speech. The model achieved high accuracy on both the training set (99.88%) and the validation set (99.29%). The convergence of loss and improvement in accuracy over the epochs indicated that the model successfully learned to classify emotions from speech data.

Additionally, exploring alternative model architectures, such as adding more LSTM layers or using bidirectional LSTMs, could potentially enhance the model's performance by capturing more complex temporal dependencies.

To apply the emotion recognition model in real-life applications, further work can focus on implementing the model for real-time predictions using streaming sensor data. This adaptation would enable the model to be deployed in real-time activity recognition systems, where it can analyze and classify emotions in real-time, leading to applications in areas such as mental health monitoring, virtual assistants, and more.

In conclusion, this study demonstrates the effectiveness of LSTM models for speech emotion recognition. The results highlight the potential of using deep learning and temporal modeling techniques to accurately classify emotions from speech cues. Future work should focus on evaluating the generalizability of the model, exploring alternative architectures, and implementing real-time applications to further enhance the practicality and usefulness of the emotion recognition system.

6. References

[1] J. Irastorz and M. I. Torres, “Analyzing the expression of annoyance during phone calls to complaint services,” in 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2016, p. 103106.

[2] S. E. Eskimez, K. Imade, N. Yang, M. Sturge-Apple, Z. Duan, and W. B. Heinzelman, “Emotion classification: How does an auto-mated system compare to naive human coders?” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016, 2016,pp. 2274–2278.

[3] A. Davletcharova, S. Sugathanb , B. Abrahamc , A. Pappachen James “Detection and Analysis of Emotion From Speech Signals” Department of Electrical & Electronic Engineering, Nazarbayev University, Kazakhstan. Enview Research & Development Labs, Trivandrum, India. Medical College Hospital, Kannur, India.

[4] Emotion Detection from Speech 007/ShahHewlett%20-%20Emotion%20Detection%20from%20Speech.pdf

[5] J. Zhao, X. Mao, L.Chen “Speech emotion recognition using deep 1D & 2D CNN LSTM networks” Biomedical Signal Processing and Control 47 (2019), pp 312-323

[6] Pandey, S. K., Shekhawat, H. and Prasanna, S. (2019). Deep learning techniques for speech emotion recognition: A review, 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), IEEE, pp. 1–6.