Speech Emotion Recognition using Deep Learning and Machine Learning Techniques

Karthini Rajasekharan
Department of Electrical and Computer Engineering
Concordia University
Montreal, Quebec
40238556

I. Introduction

This project's main objective is to create a reliable system for speech emotion classification by combining machine learning and deep learning techniques. The seven main emotions that are the focus of this study are disgust, fear, sadness, anger, happiness, surprise, and neutral. Since these feelings are essential to human expression and communication, accurately classifying them is a major task. The core of the project is manual data collecting; that is, obtaining five distinct audio signals, one for each of the following: surprise, fear, sad, neutral, and happiness. This project aims to develop a model that can accurately identify and distinguish between these various emotional states within speech signals by utilizing cutting-edge deep learning and machine learning techniques. The model may find use in sentiment analysis, customer feedback analysis, and human-computer interaction systems, among other areas.

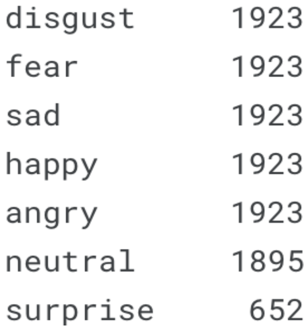
II. Previous Work

Prior studies in Human-Computer Interaction (HCI) have focused on Automatic Speech Emotion Recognition (SER), showcasing its broad applications. Key efforts include extracting speech features like Mel Frequency Cepstrum Coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC), using datasets like the Berlin emotional database for emotion classification with Support Vector Machines (SVM). Notably, LIBSVM yielded high accuracies: 93.75% overall, 94.73% for males, and 100% for females, underscoring SVM's potential in SER [1].

In earlier work by Dias Issa, M. Fatih Demirci, and Adnan Yazici, five unique features were extracted from sound files and combined into a one-dimensional array using mean values along the time axis. A 1-D Convolutional Neural Network (CNN) model utilized this array for voice emotion identification, emphasizing the importance of feature diversity for improved categorization. The researchers iteratively refined their model, likely through hyperparameter tuning, architecture adjustments, and data augmentation, enhancing its emotion recognition capabilities [2].

III. Dataset Description

The RAVDESS dataset, comprising acted speech and song performances across emotional states like neutral, calm, happy, sad, angry, fearful, disgust, and surprised, offers a diverse range of vocal expressions ideal for training emotion recognition models in speech contexts. Similarly, CREMA-D presents audiovisual recordings featuring actors portraying emotions such as anger, disgust, fear, happiness, sadness, and surprise, facilitating the development of emotion recognition systems across different modalities. TESS, with its audio recordings of sentences conveying emotional states like anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral, serves as a valuable resource for training and testing emotion recognition algorithms, particularly in speech analysis. Lastly, the Surrey Audio-Visual Expressed Emotion (SAVEE) database provides audio recordings of acted speech representing basic emotions like anger, happiness, sadness, and neutrality, voiced by multiple actors, contributing to a diverse set of vocal expressions for emotion recognition research. Together, these datasets merged offer a rich repository of **12162** emotional speech data, empowering researchers to create and assess speech-driven emotion recognition systems for various applications.



```
disgust     1923
fear        1923
sad         1923
happy       1923
angry       1923
neutral     1895
surprise     652
```

Fig. 1. Combined Dataset

IV. Audio Preprocessing

The preprocessing includes four essential audio preprocessing functions: 'noise', 'stretch', 'shift', and 'pitch'. The 'noise' function adds random noise to the input audio data, with the noise amplitude calculated as 0.035 times a random value between 0 and 1 times the maximum amplitude of the data. The 'stretch' function performs time stretching on the audio, altering its duration while maintaining pitch using Librosa's time_stretch function. The 'shift' function randomly shifts the audio data by a range determined by a random value between -5 and 5 milliseconds, simulating timing variations. Lastly, the ‘pitch‘ function modifies the audio’s pitch based on a specified pitch factor, defaulting to 0.7 if not provided, using Librosa’s pitch_shift function. These preprocessing techniques are crucial for data augmentation, enhancing the training dataset’s diversity, and improving the model’s ability to generalize and perform well across various acoustic conditions in tasks like speech recognition or emotion classification.

V. Audio Feature Extraction Method

Spectrograms and waveforms are fundamental representations of audio signals used extensively in applications like speech-driven emotion recognition. A waveform depicts the amplitude of the audio signal over time in the time domain, offering insights into temporal characteristics such as pitch, duration, and intensity variations. In contrast, a spectrogram, a frequency-domain representation obtained through the Fourier transform of small signal segments, visualizes the signal’s frequency content over time, with darker areas indicating higher energy at specific frequencies. Spectrograms provide a comprehensive view of both spectral and temporal features, making them valuable for analyzing complex audio like speech. In speech-driven emotion recognition systems, spectrograms and waveforms are key input representations for feature extraction. Spectrograms capture detailed frequency information, aiding in discerning subtle spectral patterns related to different emotions, while waveforms reveal temporal dynamics crucial for identifying emotional patterns. Leveraging both spectrograms and waveforms enables emotion recognition models to extract comprehensive features encompassing spectral and temporal cues, leading to more accurate and robust emotion classification for personalized movie recommendations.

Mel-Frequency Cepstral Coefficients (MFCCs) are a widely used feature representation technique in speech processing and emotion recognition tasks, derived from the short-time Fourier transform (STFT) of audio signals to mimic the human auditory system’s response to sounds. This involves segmenting the audio signal into overlapping frames, computing their power spectra, and applying a Mel-filterbank to transform the frequency spectrum onto the Mel scale. The resulting Mel-filterbank energies undergo a logarithm operation and then discrete cosine transform (DCT) to decorrelate the features, yielding the final set of MFCCs. These coefficients effectively capture spectral characteristics across different frequency bands, providing a compact yet informative representation. Alongside, Root Mean Square (RMS) quantifies signal energy by computing the square root of the mean squared amplitude values, useful for tasks like speech and music analysis. In emotion recognition systems, RMS values capture loudness variations aiding emotional expression discrimination. Zero Crossing Rate (ZCR), measuring how often a signal changes sign within a frame, is vital in capturing temporal dynamics, pitch modulation, and speech rhythm nuances in emotion recognition, thereby enhancing classification accuracy and system performance.

VI. Algorithms

A. Convolutional Neural Network

Convolutional Neural Networks (CNNs) are a class of deep learning models designed for processing structured grid-like data, such as images or sequential data like time series. The model generated is a professionally structured CNN built using TensorFlow’s Keras API. It begins with a 1D convolutional layer with 512 filters, employing a kernel size of 5, a stride of 1, and ”same” padding to maintain input dimensions, followed by Rectified Linear Unit (ReLU) activation for introducing non-linearity. Batch normalization is applied to standardize and stabilize the activations, enhancing training efficiency. A max-pooling layer with a pool size of 5 and stride of 2 reduces the spatial dimensions while retaining important features. This pattern repeats with additional convolutional blocks, gradually reducing the spatial dimensions with max-pooling and incorporating dropout layers (20% probability) for regularization, preventing overfitting. The final layers include a flattening operation to prepare for dense layers, which are followed by a 512-unit dense layer with ReLU activation and batch normalization for feature extraction and a 7-unit softmax output layer for multiclass classification. The model is compiled using the Adam optimizer with categorical cross-entropy loss and accuracy as the evaluation metric, ensuring efficient training and performance assessment.

B. Support Vector Machine

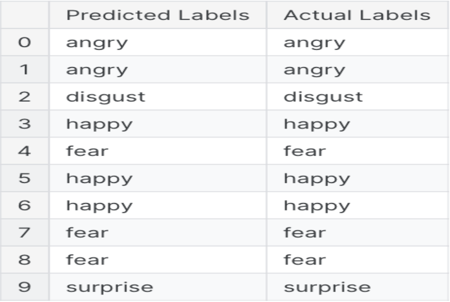
Support Vector Machines (SVM), a powerful supervised learning algorithm, for audio classification tasks. Initially, the audio data is preprocessed, extracting features such as Mel-frequency cepstral coefficients (MFCCs), zero-crossing rates (ZCRs), and root mean square energies (RMSEs) using the Librosa library. These features are essential for capturing relevant information from the audio signals. The data is then split into training and testing sets, standardized using StandardScaler to ensure consistent scaling across features, and the labels are converted from one-hot encoding to one-dimensional arrays. The SVM model is instantiated with a radial basis function (RBF) kernel, a regularization parameter (C) set to 1.0, and ’scale’ for gamma, although these hyperparameters can be fine-tuned for optimal performance. The SVM model is trained on the standardized training data, and subsequently, predictions are made on the test data to evaluate its performance. Finally, the accuracy of the SVM model is calculated using the accuracy score function from scikit-learn, providing a quantitative measure of its classification performance on the audio data.

VII. Manual Data Collection

It was a great experience when I collected the audios through microphone from different people with different emotions. I collected surprise, neutral, happy, fear and sad. Initially I faced a lot of challenges to make the people speak the way I want the emotions to be. For example, collecting the audio emotions from males was the most difficult task. Generally males are not expressive and only one person accepted to give me an audio emotion, which was happy. Rest of the audio signals were females and I collected them. Overall, I had a good experience collecting my own data.

VIII. Results

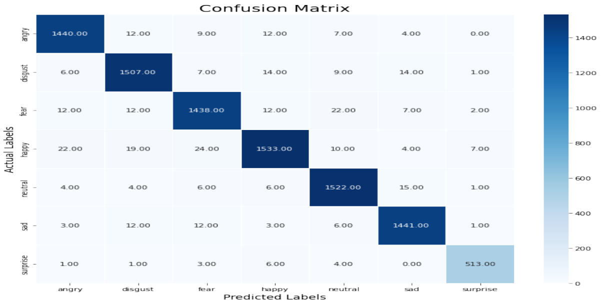
In this study, the Convolutional Neural Network (CNN) model achieved an impressive accuracy of **96.88%** in classifying voice emotions. On the other hand, the Support Vector Machine (SVM) classifier demonstrated a solid performance with an accuracy of **87.4%** in the same task. These results underscore the effectiveness of CNNs in capturing intricate patterns in voice data for emotion recognition, leading to higher classification accuracies compared to traditional SVM classifiers.



	Predicted Labels	Actual Labels
0	angry	angry
1	angry	angry
2	disgust	disgust
3	happy	happy
4	fear	fear
5	happy	happy
6	happy	happy
7	fear	fear
8	fear	fear
9	surprise	surprise

Fig. 2. Predictions of CNN on Test Data

A confusion matrix is a performance measurement tool used in classification tasks to evaluate the accuracy of a model's predictions. It presents a tabular representation of predicted versus actual classes, enabling a detailed analysis of model performance. The confusion matrix allows researchers to assess various performance metrics such as precision, recall, F1 score, and specificity, providing a comprehensive understanding of the model's strengths and weaknesses across different classes.



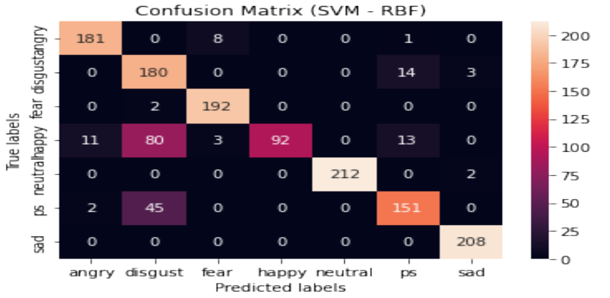
Confusion Matrix

Actual Labels \ Predicted Labels	angry	disgust	fear	happy	neutral	sad	surprise
angry	1440.00	12.00	9.00	12.00	7.00	4.00	0.00
disgust	6.00	1507.00	7.00	14.00	9.00	14.00	1.00
fear	12.00	12.00	1438.00	12.00	22.00	7.00	2.00
happy	22.00	19.00	24.00	1533.00	10.00	4.00	7.00
neutral	4.00	4.00	6.00	6.00	1522.00	15.00	1.00
sad	3.00	12.00	12.00	3.00	6.00	1441.00	1.00
surprise	1.00	1.00	3.00	6.00	4.00	0.00	513.00

Fig. 3. Confusion Matrix - CNN

The classification report is a vital tool in evaluating the performance of a machine learning model, particularly in classification tasks. It provides a detailed summary of various metrics such as precision, recall, F1-score, and support for each class in the dataset.

$$\text{Precision}(C) = \frac{\text{True Positives}(C)}{\text{True Positives}(C) + \text{False Positives}(C)}$$



Confusion Matrix (SVM - RBF)

True labels \ Predicted labels	angry	disgust	fear	happy	neutral	ps	sad
angry	181	0	8	0	0	1	0
disgust	0	180	0	0	0	14	3
fear	0	2	192	0	0	0	0
happy	11	80	3	92	0	13	0
neutral	0	0	0	0	212	0	2
ps	2	45	0	0	0	151	0
sad	0	0	0	0	0	0	208

Fig. 4. Confusion Matrix - SVM

Precision measures the accuracy of positive predictions for class C among all instances predicted as class C.

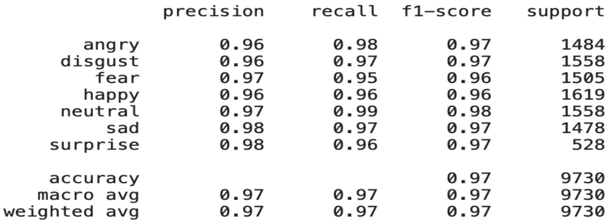
$$\text{Recall}(C) = \frac{\text{True Positives}(C)}{\text{True Positives}(C) + \text{False Negatives}(C)}$$

Recall represents the ratio of correctly predicted instances of class C to the total number of actual instances of class C.

$$\text{F1-score}(C) = 2 \times \frac{\text{Precision}(C) \times \text{Recall}(C)}{\text{Precision}(C) + \text{Recall}(C)}$$

F1-Score provides a balanced measure between precision and recall, useful for scenarios where both false positives and false negatives are important.

In a scientific project, the classification report is instrumental in assessing the model's performance across different classes, identifying areas of strengths and weaknesses, and guiding further improvements or adjustments in the model architecture or training process.

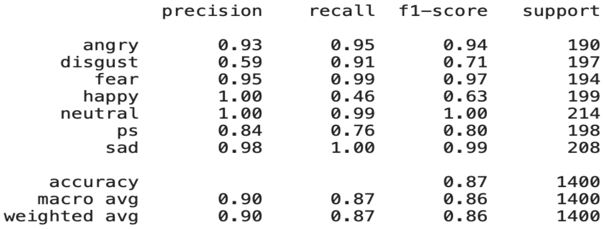


```
              precision    recall  f1-score   support

       angry       0.96      0.98      0.97      1484
     disgust       0.96      0.97      0.97      1558
        fear       0.97      0.95      0.96      1505
       happy       0.96      0.96      0.96      1619
     neutral       0.97      0.99      0.98      1558
         sad       0.98      0.97      0.97      1478
    surprise       0.98      0.96      0.97       528

    accuracy                           0.97      9730
   macro avg       0.97      0.97      0.97      9730
weighted avg       0.97      0.97      0.97      9730
```

Fig. 5. Classification Report - CNN



```
              precision    recall  f1-score   support

       angry       0.93      0.95      0.94       190
     disgust       0.59      0.91      0.71       197
        fear       0.95      0.99      0.97       194
       happy       1.00      0.46      0.63       199
     neutral       1.00      0.99      1.00       214
          ps       0.84      0.76      0.80       198
         sad       0.98      1.00      0.99       208

    accuracy                           0.87      1400
   macro avg       0.90      0.87      0.86      1400
weighted avg       0.90      0.87      0.86      1400
```

Fig. 6. Classification Report - SVM

The manually collected audio signals are predicted by both CNN and SVM and are tabulated below.

Audio File and Actual Label	CNN Predictions	SVM Predictions
AUDIO-2024-04-05-21-21-38-1 (Surprise)	Surprise	Disgust
AUDIO-2024-04-05-21-21-38.m4a (Fear)	Fear	Fear
AUDIO-2024-04-05-21-22-33.m4a (Happy)	Disgust	Happy
AUDIO-2024-04-05-21-22-47.m4a (Neutral)	Neutral	Disgust
AUDIO-2024-04-05-02-26-31.m4a (Sad)	Sad	Disgust

We have to recognize emotions from speech for various reasons,

Healthcare: Monitor the emotional well-being of individuals, especially elderly individiuals, that could help clinicians diagnose and treat mental health conditions such as depression or anxiety.

Customer Service and Support: Analyze the emotional state of callers in real time, helping customer service representatives adapt their responses to improve customer satisfaction. It can also route calls to the most appropriate agent based on the emotional context.

Team Management: Monitor the emotional state of team members during meetings or in remote work settings to better manage team dynamics and address potential issues before they escalate.

IX. Other Ideas

In this project, an alternative approach could have been to utilize videos for manual testing instead of direct audio signals. By extracting audio voice from videos automatically, the project could have expanded its scope to include additional features such as visual cues, facial expressions, and gestures, enhancing the depth of emotion recognition. Additionally, leveraging techniques to attenuate voice signals could have improved the accuracy of gender and age identification, providing richer insights into speaker demographics. Incorporating video-based analysis would have allowed for a more comprehensive understanding of emotion expression, speaker characteristics, and context, thereby enhancing the overall capabilities and versatility of the emotion recognition system.

X. Concepts Learnt

Collection of manual data was a new thing to me as I have never collected audio signals that have different emotions. I was wondering how they would fit into my two models, that is, CNN and SVM, but I made it. CNN was found to give more accurate results. This is the first time I am working with audio signals and with the motivation of our Professor Paula Lago, I completed it successfully.

XI. Project Comparison

In this section, I am going to compare my project with an another project that has used audio signals as their dataset. There is only one team that has worked on speech emotion and hence I am comparing my project with them.

The name of their project is “Post Traumatic Stress Disorder Detection Using Audio Signals”. The main goal of their project is to create a web application capable of detecting Post Traumatic Stress Disorder in the user by analyzing his/her voice. From the audio recordings, they had analyzed emotions such as anger, fear and sadness parameters with their two models such as Convolutional Neural Network and Resnet-50 and had got a result about whether the person has had Post Traumatic Stress Disorder Detection (PTSD) or not.

The main difference that we have between my project and their project is they have done an application project that includes a full-stack work while I have fully concentrated on the scientific part. Their team has

Strengths:

1. The team members have a complete knowledge about full-stack and they have built their application very well.

2. I was able to see their front-end and back-end work and I was really impressed with that.

Weakness:

1. I have used 7 different emotions while they have used only 3 emotions in their project.

2. Since they have chosen to do an application project their concentration is fully on the application side, like, to work on the front-end and back-end, rather than to concentrate on adding more models and getting appropriate accuracy for it.

3. Their pre-processing does not seem to have more steps, in cleaning the data properly. They have said “The audio is converted into MFCSS (Mel-Spectrogram) using librosa prebuilt library in Python. Then those melspectrograms were passed through in a pre-trained Resnet-50 model through which we were able to train our model and with the help of FastAI and Pickle we were able to save the model and load it any flask model.” According to my project, I have completed Data Augmentation, where I used noise, stretch, pitch and shift to clean the data.

4. Accuracy of the model is not predicted. They have obtained a probability of success given by the algorithm that they have used to train their model (Resnet-50). It was around 93% and they were getting a ”high” value because their dataset was very small.

References

[1] Y. Chavhan, M. L. Dhore, and P. Yesaware, “Speech emotion recognition using support vector machine,” *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[2] D. Issa, M. F. Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” *Biomedical Signal Processing and Control*, vol. 59, p. 101894, 2020.