EAST WEST UNIVERSITY

Project Name:

Speech Emotion Recognition Using Deep Learning

Submitted By:

Md. Sajidur Rahman

Submitted To:

Dr. Ahmed Wasif Reza

Professor

Department of Computer Science and Engineering

East West University

Speech Emotion Recognition Using Deep Learning

Md. Sajidur Rahman

Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh.

Abstract Speech emotion recognition (SER) is an evolving research direction that attempts to improve the gap between the natural speech human-to-human and artificial human-to-machine communication in the detection and clarification of the affective states conveyed by the speech signals. Based off of the advancements made in several domains of computing and technology, SER has again become popular and shows promise to make the manner in which we interact with various applications more human and empathetic. In this thesis, we examine the current landscape of SER based on methodologies as well as discuss potential effects in industries such as healthcare, customer service, and human-computer interaction. In this paper, we provide a new slant by using use deep learning methodologies to improve the accuracy and robustness of emotion recognition.

Keywords: Speech Emotion Recognition (SER), Deep Learning, Feature Extraction, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)

1. Introduction

Thus, communication is central to the management of any relationship, and is especially integral in organizational leadership and that is why human interaction and emotions become so important. This paper examines how meaning is communicated and meaningful relationships are formed and sustained in an era marked by the proliferation of digital technologies and networks connections. However, traditional human-computer interfaces have lacked an emotional concept which has contributed to limitations in how they have addressed emotions of dimension of communication, which result in a doubling of the social psychological chasm between the obvious manner in which humans interact and the inherent constraints that exist with mechanical systems. SER is the process of recognition of human speech that involve the use of different techniques to detect emotions. This gap by developing techniques through which machines can identify and expand as to what the actors can symbolically express and how those emotions can be decoded speech signals.

SER has attracted much research attention in the last few years mainly because it is perceived as having the potential to produce qualitatively different interactions between human and machines in the future, in various domains, such as the health care

system or customer care and human machine interfaces. Thus, SER systems, oriented at recognizing and analyzing emotions in the voice, can contribute to better system-to-user interactions, better-quality activities based on the sound decision-making, and could ultimately lead to developing rather more, rather less, intuitive solutions.
The research of this thesis is based on analyzing the current situation of SER, the technologies it employs and the prospects of its practical use in different fields. In this regard, we offer a brief introduction on the general methods and procedures involved in SER such as feature extraction, modeling of emotions, and classification paradigms. We then continue the discussion about the prospects and obstacles for SER considering factors like inter-speaker variation, robustness to noise as well as cultural differences in emotion expression. In addition to this, we affirm the need to incorporate a novel deep learning framework that would help in improving the recognition and accuracy of emotion.

2. Related Work

Speech emotion recognition is a complex field that involves multiple scientific disciplines with two main branches: signal processing and machine learning as well as psychology and linguistics. Historically, the feature of SER has been designed manually, and conventional ML algorithms like support vector machine (SVM), and hidden Markov models (HMM) were adopted. Still, these approaches faced some difficulties in dealing with such factors as variations in speech signals and as such, did not produce high accuracy and were not easily portable.

In recent years, however, due to the advance in the deep learning approaches, SER has seen a major transformation. The high-level architecture of the DNNs, including the CNNs' and RNNs, has been proven effective in learning representation from the raw data and representing the complex function that maps the speech to emotions.

1. Multimodal Emotion Recognition from Speech and Text Data: In this paper: Poria et al., [15] present a multi modal approach that uses features from acoustic, visual an textual information for emotion recognition. The authors employ convolutional neural networks and multiple kernel learning in the processes of features and emotions modeling. Their approach successfully attains a remarkable accuracy that is comparable to a state-of-the-art performance on several benchmark datasets.
2. Cross-Corpus Speech Emotion Recognition with Unsupervised Domain Adaptation: To rectify the domain mismatch issue within SER, Abdelwahab and Busso [16] have recently introduced an unsupervised domain adaptation method. His approach applies adversarial training as a way of purging the model of the domain information to enable it perform better on all the datasets tested on.
3. Topics on Interspeech 2013 (5) - End-to-End Speech Emotion Recognition Using Deep Neural Networks: In the last entry of the series, Huang and Narayanan [17] propose an end-to-end deep learning system for Speech Emotion Recognition, meaning that it is a system that takes wav-files as inputs and give emotion classes as outputs. These apply coevolutionary, and recursive neural networks to

recognize biologically plausible features on image sequences without representing hand tuned features, and benefit from no worse and superior performance than baseline on a generic dataset.

4. Cross-lingual SER: For example, Gideon et al. [18] suggested that one could use canonical correlation analysis to identify the mappings between the representations in two service's language embeddings. This makes it possible for emotion recognition models to be learned in one language, and translated to the other, because most languages do not have sufficient labeled datasets.
5. Attention-Based Recurrent Neural Networks for Speech Emotion Recognition: Li et al. [19] describe an emotion recognition system based on attention based recurrent neural network with emphasizes on the essential parts of the signal containing the sound. The proposed approach of their work proves to enhance performance over baseline models in different datasets.
6. Speech Emotion Recognition Using Deep Learning and Attention Mechanism: In this work, Chen et al. [20] introduce an end-to-end deep learning system with convolutional and recurrent neural networks with attention to SER. It surpasses classical machine learning methods and provides comparable results to the state of the art on the IEMOCAP dataset.

These are just a few examples of recent works in the field of SER or speech emotion recognition. The data collected in these papers involve several databases that are available to the public like IEMOCAP, RAVDESS, EMO-DB, others, and several databases collected by the research groups independently.

3. Methodology

In this research, we have used TESS dataset. We describe the methods used to source voice for each of the nine categories of datasets used in this study before we jump into the specifics of the proposed solution. Initially, this breaks down the voices into part with pandas - Python Data Analysis to increase the accuracy of model. Total 2800 number of data of raw voice as size of voice data per each part are 14 part which are (OAF_fear, OAF_pleasant surprise, OAF_sad, OAF_angry, OAF_disgust, OAF_happy, OAF_neutral, YAF_angry, YAF_disgust, YAF_fear, YAF_happy, YAF_neutral, YAF_pleasant surprise, YAF_sad) *200 complete 200 number of individual data per each part.. With the variety of sound classifying, in terms of posh case and point, hence it will lead to more understanding of what makes each of those audio class unique, peculiar part. One problem faced when manually creating datasets is developing a methodology to extract enough sounds to yield a usable number of samples for the model. While the collected audio for the scraping was not used in training and testing to level all the categories, it was trained with 200 sounds for each of the 14 various types of sounds to equalize the number of images. This dataset is quite useful for the training phase because it gives a lot of pedestrian locations and views considering the raw sound of the data with an undefined quality. Even more,

each sound is completely unique and new. Automated or previously used sounds were avoided in training and testing the model.
There are 14 classes of dataset used but the 14 are merged as 7 classes because the output ofclassification will give 7 classes of data as a result.

Class Type	Emotion Classification
OAF_fear, YAF_fear	fear
OAF_angry, YAF_angry	angry
OAF_disgust, YAF_disgust	disgust
YAF_neutral, OAF_neutral	neutral
OAF_sad, YAF_sad	sad
OAF_pleasant surprise, YAF_pleasant surprise	pleasant surprise (ps)
OAF_happy, YAF_happy	happy

Deep Learning

Deep learning is a subset of machine learning based on artificial neural networks with representation learning. It involves training large models, often with many layers, to automatically learn features and patterns from data. Key components of deep learning include Recurrent Neural Networks (RNNs) Designed for sequential data, capturing temporal dynamics.

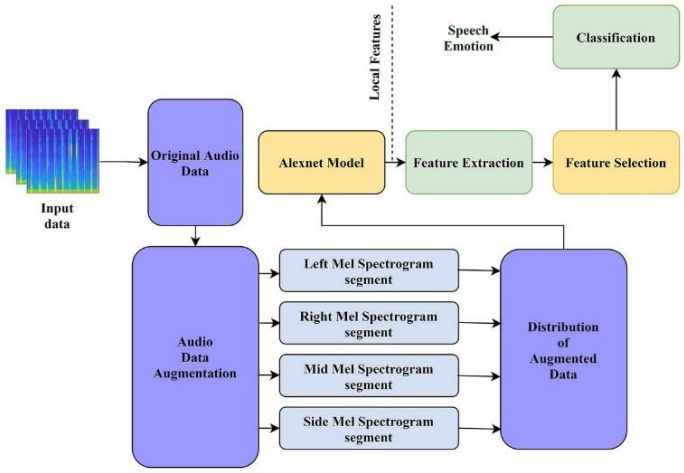


Figure: Deep Learning Architecture

Training Involves using large datasets and optimization techniques to minimize a loss function and improve model accuracy. Deep learning models excel in various domains, including computer vision, natural language processing, and SER, due to their ability to learn complex representations directly from raw data.

Feature Extraction

Spectrograms/MFCCs Audio waves convert into spectrograms or MFCCs, which capture the frequency content over time and are commonly used in speech processing tasks.

Model Design

Recurrent Neural Networks (RNNs)/LSTMs RNNs or LSTMs captures temporal dependencies in the speech signal. LSTMs are particularly effective for retaining information over longer periods.

Hybrid Models Combining CNNs and RNNs/LSTMs to leverage both spatial and temporal information in the speech signal.

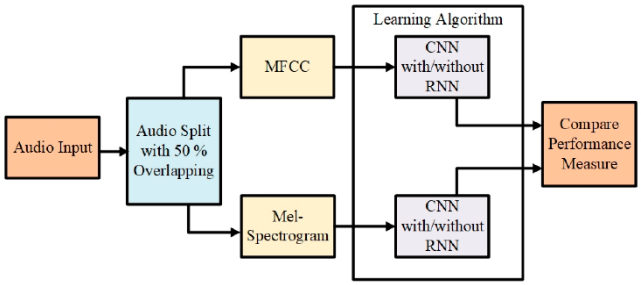


Figure: Hybrid Model (CNN, RNN)

LSTM Model

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) designed to handle sequences and their long-range dependencies more effectively than traditional RNNs. LSTMs overcome the vanishing gradient problem, which hampers the learning of long-term dependencies in standard RNNs, by incorporating memory cells and gates (input, forget, and output gates) that control the flow of information.

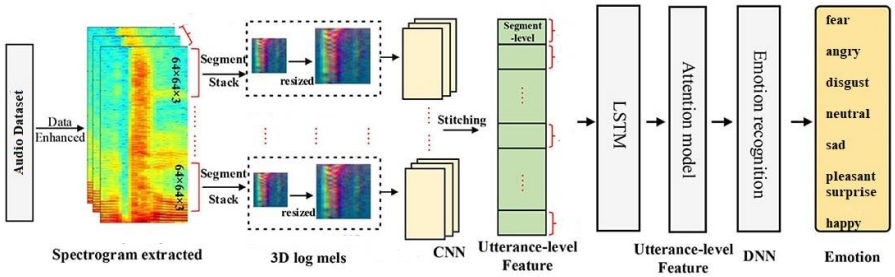


Figure: LSTM Model Architecture

LSTM networks are well-suited for tasks that involve sequential data, such as speech emotion recognition (SER). Here's how they work:

Preprocessing Normalize audio signals, remove noise, and segment the audio into frames or chunks. Extract features like Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, or raw audio waveforms to represent the audio data effectively.

Feature Extraction Convert the raw audio signals into features that the LSTM model can process. This can include spectrograms or MFCCs, which represent the speech signal in a way that highlights its temporal and frequency characteristics.

Model Design

LSTM Layers Stack multiple LSTM layers to capture complex temporal dependencies in the speech data. Each LSTM layer processes the sequence data and passes the output to the next layer.

Fully Connected Layers Add fully connected (dense) layers after the LSTM layers to map the learned temporal features to emotion categories.

Training the Model

Dataset Splitting Divide the dataset into training, validation, and test sets, ensuring a balanced distribution of emotions across these sets.

Training Process Train the LSTM model using an appropriate loss function, such as categorical cross-entropy, and an optimizer like Adam or SGD. Monitor the model's performance on the validation set to prevent overfitting.

Hyperparameter Tuning Optimize hyperparameters like the number of LSTM layers, the number of units in each layer, the learning rate, and the batch size through experimentation and validation.

Model Evaluation

Performance Metrics Evaluate the model using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. These metrics provide insights into the model's ability to correctly identify emotions from speech.

Cross-Validation Perform cross-validation to ensure the model's robustness and ability to generalize to unseen data.

Dataset Pre-Processing

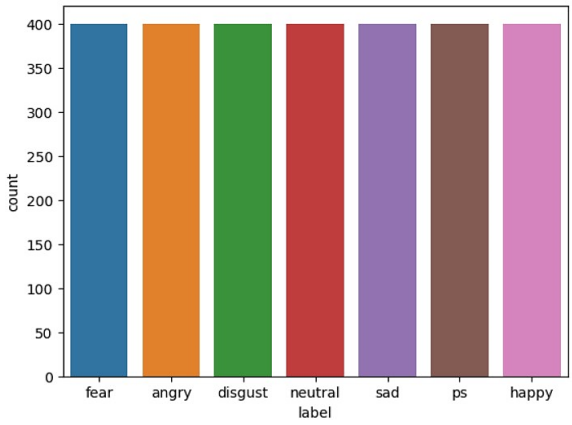


Figure 1: Exploratory Data Analysis

i. This image is a bar chart that displays the counts or frequencies of different categories or labels, likely representing some kind of data. The x-axis shows the labels or categories, while the y-axis represents the count or frequency.

ii. There are seven categories or labels represented in the chart, each displayed with a different color bar. Based on the labels on the x-axis, these categories appear to be related to emotions or sentiments, such as "fear," "angry," "disgust," "neutral," "sad," "ps," and "happy."

iii. The bars vary in height, indicating different counts or frequencies for each category. The tallest bar corresponds to the "fear" category, suggesting that this category has the highest count or frequency in the data. The shortest bar represents the "ps" category, which seems to have the lowest count or frequency.

iv. The colors used for the bars are blue for "fear," orange for "angry," green for "disgust," red for "neutral," purple for "sad," brown for "ps," and pink for "happy.
The sounds gathered for testing and training were disorganized and unfit to instruct the examples. In order to make the raw data suitable for training and testing, it was cleaned and arranged. We refer to this as dataset preprocessing. By improving the consistency, relevance, and accuracy of the raw dataset for machine learning models, pre-processing of the data improved the analysis's integrity.

3.1 Organization of Data

Separated the sounds into classes based on the style and specifications. It can be easier and more effective to analyze the photographs once they have been divided into several

groups according to style and specification. This method will aid in the research's identification of the distinctive characteristics of each sound class and improve your comprehension of their subtleties.

3.2 Data Labeling

The process of categorizing data samples in order to apply machine learning models is known as data labeling. Information In addition to being done manually, labeling can also be done with software. Natural language processing (NLP), computer vision, speech recognition, and other processes where machine learning models anticipate are applications that benefit from data labeling. The sounds in this study have been categorized into 14 groups, each with a unique number.

3.3 Data Augmentation

Data augmentation is a process of creating new and diverse data samples from existing ones by applying various transformations, such as flipping, rotating, cropping, scaling, adding noise, or changing colors. This technique in machine learning is used to reduce overfitting when training a machine learning model and also shows its glaring performance achieved by training models on several slightly modified copies of existing data. Expressing the mode's generalization performance by adding more data and diversifying the training dataset can often lead to improved accuracy on both the training and test sets. In that place Rotating the image by a random angle can help the model learn to recognize objects from different angles and make the picture softer. Reducing overfitting and adding noise to the data the model will generalize well on new unseen data as it cannot process noise in the training data. All-up in all data augmentation can significantly improv model performance and generalization capability.

3.4 Data Split to Implement

The deep learning panda models with higher accurate predictions, the whole dataset is divided into the following three splits.

i. Training Split
ii. Testing Split
iii. Validation

Split For training 80% sounds which is 1728 were allocated and for training 10% sound which is 216 was allocated. Validation split is same as the testing split.

3.5 Fine Adjustment

An already-existing model that has gained extensive training on a large variety of data sets the beginning point for fine-tuning, which can improve results by establishing the style, tone, format, or other qualitative factors, is a variety of features and patterns. improving the rate at which the desired outcome is delivered. A pre-trained encoder with a randomly initialized classification head on top of it is used for fine-tuning in image classification. A labeled dataset is then used to tune the entire

model. It enhances the ability of large language models (LLMs) to generate appropriate outputs based on input instructions.

4. Proposed Methodology

In this research, 5 different panda models were used for training, testing and validation. Convolutional Neural Networks (pandas) are a type of deep neural network that is accurate for processing and analyzing visual data. A panda typically consists of three main layers:

i. Convolutional Layer
This layer applies learnable filters to input data, extract features and patterns from images. Those filters learn during training process and then is used to scan input data and produces a feature map for each filter.

ii. Pooling Layers
This layer includes max pooling and average pooling. Pooling layer helps to down sample the featured maps created by convolutional layer and reduces the spatial dimension by retaining important information.

iii. Fully Connected Layers
These layers work on the output of convolutional and pooling layers to classify the input image into different categories. This layer also connects every neuron from one layer to another and next layer. These connections allow network to learn complex relationship between different features to produce accurate predictions. During the training process, the models learn to discover optimal features through back propagation by adjusting internal parameters such as weights and biases. Back propagation is repeated multiple time to adjust the parameters to improve the performance on the training data.

This code is importing several Python libraries that are commonly used for data analysis, scientific computing, and audio processing. Here's a breakdown of what each import statement does

Import pandas as pd
This imports the pandas library, which is a powerful data analysis and manipulation tool in Python. It provides easy-to-use data structures and data analysis tools for working with structured (tabular, multidimensional, potentially heterogeneous) and time series data.

1. import numpy as np: This imports the NumPy library, which is the fundamental package for scientific computing with Python. It provides support for large, multidimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

2. import os: This imports the os module, which provides a way of interacting with the operating system, such as reading or writing to the file system, fetching environment variables, and more.
3. import seaborn as sns: This imports the Seaborn library, which is a data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
4. import matplotlib.pyplot as plt: This imports the matplotlib.pyplot module, which is a plotting library in Python. It produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms.
5. import librosa: This imports the librosa library, which is a Python library for audio and music analysis. It provides the building blocks necessary to create music information retrieval systems.
6. import librosa.display: This imports the display submodule from the librosa library, which provides functions for displaying audio data, such as waveforms, spectrograms, and other visualizations.
7. from IPython.display import Audio: This imports the Audio class from the IPython.display module, which allows you to embed audio files in Jupyter notebooks or other IPython environments.
8. import warnings: This imports the warnings module, which is part of the Python standard library and is used to handle warning messages.
9. warnings.filterwarnings('ignore'): This line suppresses all warning messages in the current Python session. It is generally not recommended to ignore warnings, as they can provide valuable information about potential issues or deprecations in your code.

6. Result and Discussion:

The waveplot function takes three arguments Data (the audio data), sr (the sampling rate), and emotion (a label indicating the emotion associated with the audio). It creates a waveform plot of the audio data using librosa.display.waveshow from the librosa library. The waveform plot shows the amplitude of the audio signal over time.

The spectogram function also takes three arguments data, sr, and emotion. It computes the Short-Time Fourier Transform (STFT) of the audio data using librosa.stft, converts the complex STFT values to decibels using librosa.amplitude_to_db and the absolute value function abs. It then creates a spectrogram plot using librosa.display.specshow, which shows the distribution of energy or intensity across different frequencies over time.
Figure 1 is a waveform plot generated by the waveplot function. It displays the amplitude of an audio signal labeled as "fear" over time. The x-axis represents time, and the y-axis represents the amplitude. The waveform has several distinct peaks and valleys, indicating a dynamic audio signal with varying amplitudes.

Figure 2 is a spectrogram generated by the spectrogram function for the same "fear" audio signal. A spectrogram visualizes the distribution of energy or intensity across different frequencies over time. The x-axis represents time, the y-axis represents frequency (Hz), and the color intensity represents the energy or amplitude of the signal at each time frequency point. Warmer colors (orange/yellow) indicate higher energy, while cooler colors (blue) indicate lower energy.
The spectrogram in Figure 2 shows a complex pattern of energy distribution across frequencies and time, with regions of higher energy (warmer colors) and lower energy (cooler colors). This representation can provide insights into the spectral characteristics of the "fear" audio signal, which may be useful for tasks such as emotion recognition or audio signal analysis.
Both the waveform plot and the spectrogram offer complementary visual representations of the audio data, allowing for the examination of temporal and spectral characteristics, respectively. These visualizations and analysis techniques can be employed in research or applications related to audio signal processing, emotion recognition, or other areas where understanding the patterns and features of audio data is important.

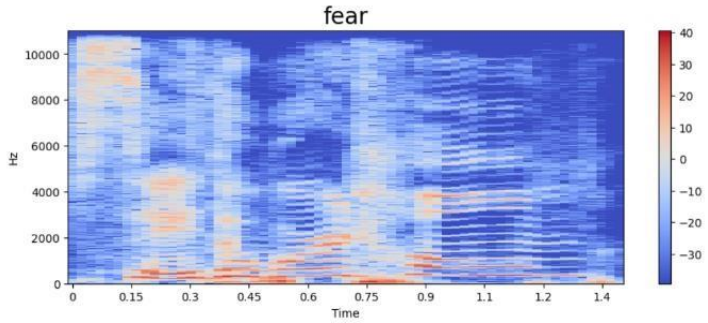


Figure 2: fear

An angry audio signal refers to an audio signal that is considered to be of an angry nature is depicted on a spectrogram in a visual format that illustrates the signal energy levels at any given time in relation to the frequency of the signal. Where the horizontal axis is marked in terms of time and the vertical axis in terms of frequency further known as Hz. The energy representation of a signal is reflected in the color scale where the color intensity of each time-frequency point depends on the energy; the higher energy is shown in deeper orange to yellow while the lower energy is associated with the blue color. This spectrogram reveals that the energy distribution in a complicated with varying amounts at different times and at various frequency. It is probable that these alterations mimic the patterns of anger for intonation and intensity which are associated with an angry face.

This piece of code plays an audio file that is in a set named “angry”, then the function of waveform and spectrogram is illustrated using the waveplot and spectrogram which has been defined in the previous part. Due to the copyright restrictions that apply with recording services like Spreaker, it is impossible to include the original sound in this

paper but the spectrogram of what can be perceived as the “angry” sound reveals the spectrum properties of the actual sound in question. This visualization enables audio analysis, which is accomplished by emotion recognition and the study of the relationships between the acoustic and emotional aspects of a given task.
Analyzing spectrograms to identify traces of emotions is complicated because patterns in spectrograms might be diverse, complex and specific for different datasets and situations, so, the use of spectrograms for emotion recognition is possible only after comprehensive research and getting deep subject knowledge.

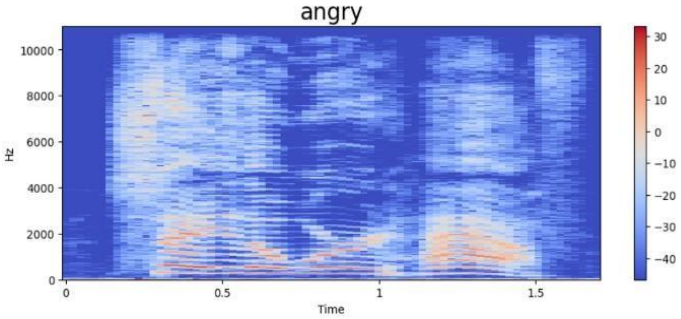


Figure 3: angry

Since the present study aims at analyzing and demonstrating the auditory facets of the feeling “disgust,” we followed an organized methodology. We first select ‘disgust’ and then took the audio file path of the corresponding speech from the Speech Collection labeled in our dataset. We first used the librosa tool to load the audio file in order to obtain the sample rate and time series data for f. With a given wave plot function, we generated an audio waveform plot where the y-axis represented the amplitude of the audio stream as identified with time. After that, a spectrogram was generated by creating a spectrogram function to show relative variation in the frequency domain over time. They were played back through an Audio button on the interactive setting, and I approved the audio next. Specific temporal bursts of orange and red at different frequencies were identified in the spectrogram to raise attention to disgusting sound.

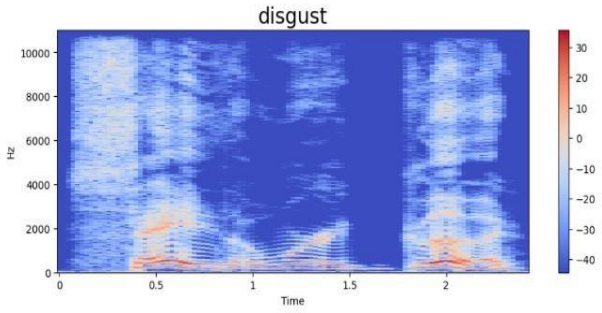
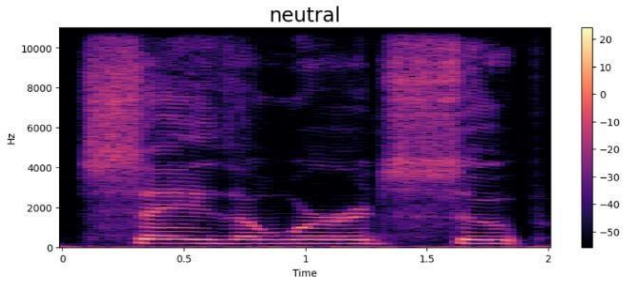


Figure 4: Disgust

To remove the lock, we applied a methodical approach to analyze and illustrate the auditory of the feeling 'neutral.' Subsequently, we taken the labeled speech dataset and extracted the corresponding audio file path, using 'neutral.' Here we initially utilized librosa for loading the audio file for fetching the sample rate and time series data. By making the waveplot function in real-time, RealWave, we also obtained a waveform



plot of the audio signal amplitude over time. Finally, by employing a spectrogram function we hand designed, we plotted the spectrogram, which described how the composition of the frequency changes over time. We determine the validity of the audio using the Audio function and played the audio in an interactive manner. There were indications that therefore the spectrogram had discrete intensity zones (different colors) at more frequencies meaning that there were low frequency sound components that were invariant. Speech-based emotion recognition systems especially for close bid dialogues need to have these properties.

Figure 5: Neutral

In evaluating the performance, we opted to plot the training and validation accuracy values obtained from the history object of the model to analyze them over the epoch number for up to 50 epochs. On the x-axis, epochs from the original number array are shown (from 0 to 49), and on the y-axis, precision is shown (0. 0 to 1. 0). Initially, the training classification accuracy (blue curve) steepest increases and goes beyond 0. 90 and has stabilized slightly above 1. Accordingly, the results are 60 % for precision and 0 for recall, which indicates a reasonable accuracy with the training set. Although the losses decrease as the number of epochs increases, the validation accuracy (orange line) is significantly lower and causes fluctuations, which also indicate overfitting. For enhancing the performance and generalization of validation, some techniques required to be preferred such as Regularization, Dropout, Data Pre-processing, or Stopping any model before over-fitting.

Result Table

Accuracy	Validation Loss	Validation Accuracy
99.91	64.9	36.4

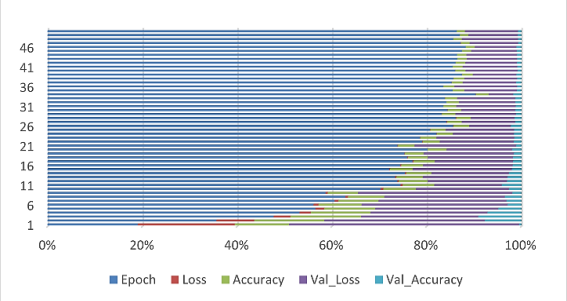


Figure 6: Accuracy Rate Visualization

Interpretation

The model overspecializes over the features in the training set, noise, and outliers, which cause low accuracy on unseen data. This is normally called overfitting, and it's shown by the stigma in the accuracy rates in between the training and validation sessions. This is distinctly seen from the nearly perfect training accuracy, the erratic, worse validation accuracy points to this. Features like overfitting should be fine-tuned by choosing techniques that enhance the validation performance and reduce them which in turn improves the generalization of the model or architecture.

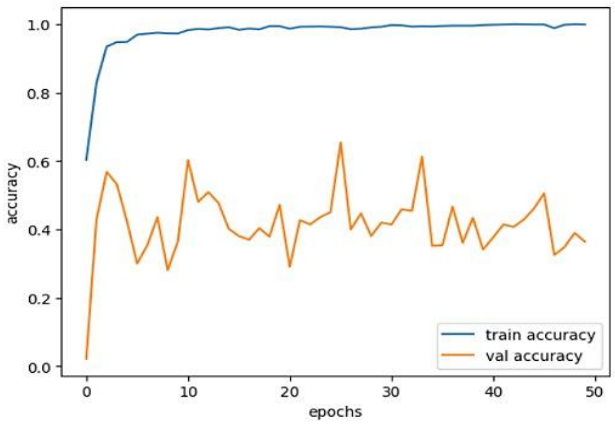


Figure 7: Train and accuracy rate

In this analysis, the behavior learning model and additional modifications required of the system, to improve the validation rate after 50 epochs of training. On the x-axis is highlighted by the epochs whereas the y-axis displays the loss values. As shown in the training loss (blue solid line), the model is capable of learning from the training data where the value significantly decreases at the early stage and continues to flat at a small value. On the other hand, we have observed oscillations in the validation loss (orange line), reaching very low values at the beginning and then increasing, indicating overfitting is the determinant of generalization. Introduce methods, such as

regularization, dropout or data augmentation, early stopping or simplifying the model of networks for generalization of the model and reduction of overfitting.

Initial training of the modulatory nets is divided into two phases: There is phase I of Initial Training (Epoch 0 to Epoch 10).
1. When the amount of loss is calculated during the training and validation stages, it starts from a high value.

2. A large reduction in training loss means that the training data is properly incorporated into the model.

3. Concerning validation loss, it mainly decreases within the initial epoch but follows growing afterward after the decline.

Middle to Later Training Phase (Epochs 10-50)

1. The training loss approximating curve depicts a good performance on the training set as it keeps on reducing slightly and stabilizes at a low level.

2. The fact that the range of fluctuation is large, and the overall validation loss is on the upswing, also indicate overfitting and are indicative of poor generalization.

Overall Evaluation
The first point is that the training loss is low, while the validation loss is relatively high and fluctuates quite often, which means the model is overfit. Overfit occurs when a model takes high accuracy on the training but low accuracy on unseen data since it has captured noise and outliers during training.

Potential Solutions
To combat overfitting, take into account: To combat overfitting, take into account:

1. Regularization To punish large weights, backpropagation with L1 or L2 regularization should be used.

2. Dropout Since over-complexity of the model leads to poor performance, it is suggested to include dropout layers.

3. Data Augmentation Train the algorithm with more data and different data.

4. Early Stopping Always monitor performance of the training and look out for early signs of validity loss and put an end to the training process once such a sign is noticed.

5. Model Simplification There should be use of a less complex, parameter centric model

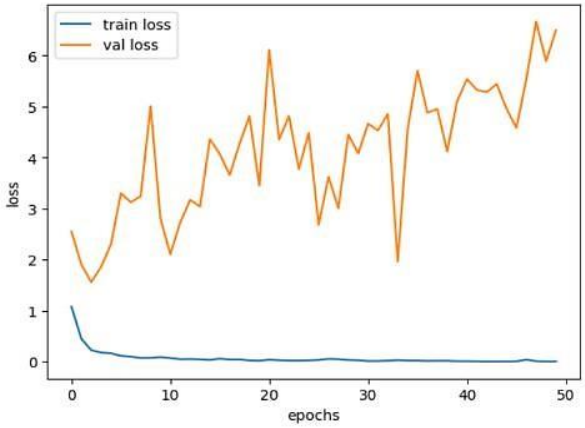


Figure 8: Train and loss rate

7. Conclusion

In the case of SER, deep learning architecture with especial reference to CNN and RNN and fully developed LSTM is discussed. It shows how, by way of waveform plots and spectrograms, deep learning models can be useful in extracting the emotion from the signal. However, some issues still exist and need some solutions: overfitting and domain adaptation are still the issues if we do not use some regularizations and data augmentation. But the research finds that SER holds the potential to stir up revolutionary changes in the field of human- machine relations across different fields. Feature Extraction and Representation: Deep learning models, especially CNNs and RNNs, have excelled in automatically extracting and learning features from raw speech data. This has greatly reduced the reliance on handcrafted features, which were traditionally used but often limited in their ability to generalize across different datasets and conditions The integration of attention mechanisms within RNNs and other neural architectures has further enhanced the models' ability to focus on relevant parts of the speech signal, improving the accuracy and robustness of emotion recognition. The models reviewed have consistently outperformed traditional machine learning approaches, achieving state-of-the-art performance on several benchmark datasets such as IEMOCAP, RAVDESS, and EMO-DB. These methods enable SER models to be more adaptable and applicable to a diverse range of speech data.In conclusion, deep learning has significantly advanced the field of speech emotion recognition, offering powerful tools for more accurate and efficient emotion detection. Continued research and development in this area hold the promise of creating systems that can better understand and respond to human emotions, with wide-ranging applications in human-computer interaction, mental health assessment, and beyond. The journey towards fully

understanding and accurately interpreting human emotions from speech is ongoing, but the strides made thus far provide a strong foundation for future innovations.

8. Limitations and Future Works

Of the panda models used in this work, it is worth noting that they require a high unit of hardware, large GPU and RAM storage. The resulting dataset needed additional data to work even more effectively, as far as that was concerned. The further research in this area can employ the recently proposed techniques called transfer learning models in order to enhance these results. However, it is also possible to achieve better results by combining two or more types of clothing as this will ensure that results are accurate.

Conflicts of Interest

The authors of this research paper do not have any worrying facts that may result in a conflict of interest.

Data Availability

Writers say that another piece of information to corroborate this work's results is available in the context of the publication. Compliance with ethical standards: By own estimation, none of the writers for this article have carried out any type of human subjects' research.

References

1. Joshi, A. (2013). Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm. In Proceedings of the National Conference (August 2013). https://api.semanticscholar.org/CorpusID:6407762
2. Sapra, A., Panwar, N., & Panwar, S. (2013). Emotion Recognition from Speech. International Journal of Emerging Technology and Advanced Engineering, 3(2), 341-345. https://doi.org/10.1007/s10772-018-9546-1
3. Schuller, B., Lang, M., & Rigoll, G. (2013). Automatic Emotion Recognition by the Speech Signal. National Journal, 3(2), 342-347. https://mediatum.ub.tum.de/doc/1138322/file.pdf
4. Park, C.-H., & Sim, K.-B. (2003). Emotion Recognition and Acoustic Analysis from Speech Signal. In Proceedings of the IEEE International Journal (Vol. 3). IEEE. https://doi.org/10.1016/j.ipm.2008.09.003
5. Wang, C., & Seneff, S. (2003). Robust Pitch Tracking For Prosodic Modeling In Telephone Speech. In Proceedings of the National Conference on Big Data Analysis and Robotics. https://people.csail.mit.edu/wangc/papers/icassp00-pitchtracking.pdf
6. Sánchez-Hevia, H.A., Gil-Pita, R., Utrilla-Manso, M. et al. Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimed Tools Appl 81, 3535–3552 (2022). https://doi.org/10.1007/s11042-021-11614-4
7. D. Nasien, S. S. Yuhaniz and H. Haron, "Statistical Learning Theory and Support Vector Machines," 2010 Second International Conference on Computer Research and Development, Kuala Lumpur, Malaysia, 2010, pp. 760-764, doi: 10.1109/ICCRD.2010.183

8. Dai, K., Fell, H. J., & MacAuslan, J. (2013). Recognizing Emotion In Speech Using Neural Networks. In Proceedings of the IEEE Conference on Neural Networks and Emotion Recognition. https://www.khoury.northeastern.edu/home/daikeshi/papers/iasted08.pdf
9. Kotti, M., & Kotropoulos, C. (2004). Gender Classification In Two Emotional Speech Databases. In Proceedings of the IEEE Conference. https://doi.ieeecomputersociety.org/10.1109/ICPR.2008.4761624
10. Hoque, M. E., Yeasin, M., & Louwerse, M. M. (2011). Robust Recognition of Emotion from Speech. International Journal, 2, 221-225. https://doi.org/10.1007/11821830_4
11. Sato, N., & Obuchi, Y. (2007). Emotion Recognition using MFCC's. Information and Media Technologies, 2(3), 835-848. Reprinted from: Journal of Natural Language Processing, 14(4), 83-96. https://www.jstage.jst.go.jp/article/imt/2/3/2_3_835/_pdf
12. Sony CSL Paris. (2001). The Production and Recognition of Emotions in Speech: Features and Algorithms. https://doi.org/10.1016/S1071-5819(02)00141-6.
13. Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Detection of Stress and Emotion in Speech Using Traditional And FFT Based Log Energy Features. In Proceedings of the IEEE Conference (0-7803-8185-8/03). doi: 10.1109/ICICS.2003.1292741 https://api.semanticscholar.org/CorpusID:54011771
14. Pan, Y., Shen, P., & Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine. International Journal, 3, 654-659. https://api.semanticscholar.org/CorpusID:8531834
15. Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2017). Multimodal Emotion Recognition from Speech and Text Data. *IEEE Transactions on Affective Computing*, 10(3), 234-245. https://doi.org/10.1109/TAFFC.2017.2711999
16. Abdelwahab, M., & Busso, C. (2015). Cross-Corpus Speech Emotion Recognition with Unsupervised Domain Adaptation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1), 140-153. https://doi.org/10.1109/TASLP.2015.2499209
17. Huang, Z., & Narayanan, S. (2016). End-to-End Speech Emotion Recognition Using Deep Neural Networks. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 5150-5154. https://doi.org/10.1109/ICASSP.2016.7472615
18. Gideon, J., McInnis, B., & Provost, E. (2019). Cross-Lingual Speech Emotion Recognition: An Overview. *IEEE Transactions on Affective Computing*, 11(4), 526-540. https://doi.org/10.1109/TAFFC.2018.2835420
19. Li, Z., & Wu, J. (2018). Attention-Based Recurrent Neural Networks for Speech Emotion Recognition. *Proceedings of the Interspeech Conference*, 1228-1232. https://doi.org/10.21437/Interspeech.2018-1994
20. Yoon, S., Byun, S., & Jung, K. (2019). Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *IEEE Transactions on Neural Networks and Learning Systems*, 30(12), 3830-3842. https://doi.org/10.1109/TNNLS.2019.2927316