IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Speech Audio Emotion Detection Using LSTM

[1]Asjad Moiz Khan, [2]Syed Mohammad Moiez, [3]Sandeep Kumar

[1]UG Student, [2]UG Student, [3]Associate Professor

[1,2,3]Department of CSE,

[1,2,3]Sharda University, Greater Noida, India

Abstract: This study explores emotion recognition in speech using Long Short-Term Memory (LSTM) networks on the Toronto Emotional Speech Set (TESS) dataset. Mel-frequency cepstral coefficients (MFCC) features are extracted, and an LSTM model is trained with a specific architecture, including Rectified Linear Unit (ReLU) and Linear activation functions in dense layers to get the membership of an audio sample in all the 7 labels and also compare them.The compiled model, utilizing mean squared error as the loss function and the Adam optimizer, achieves a training accuracy of 99% and a validation accuracy of 100%. Notably, our approach includes feature membership analysis, providing insights into the contribution of individual features to each emotion label.This comprehensive analysis demonstrates the efficacy of the proposed LSTM architecture in accurately identifying emotional states in speech, while the feature membership analysis adds a nuanced layer of interpretability to the model's predictions. The success of our approach suggests potential applications in emotion-aware technologies and interactive systems.

Index Terms **- Speech Emotion Recognition, Emotive Speech dataset, Audio data Visualization, MelFrequency Cepstral Coefficients, Feature Extraction, Long-Short Term Memory Neural Network, Emotion Classification, One Hot Encoding, Loss and accuracy matrices, Prediction Probabilities, Emotion identification, Virtual Assistants, Mental Health Screening, Computer-Human interaction**

I. INTRODUCTION

Speech Emotion Recognition (SER) has arisen as a blooming topic with extensive and diversified ramifications in an era where human-computer interaction and comprehension of human emotions play crucial roles in technology and its applications. SER aims to bridge the gap between human emotional expression through speech and algorithmic analysis of such feelings. This cutting-edge technology has the ability to improve virtual assistants, improve mental health screening, change customer service, and have an impact on a variety of other fields.The offered code highlights the SER technique, providing a thorough examination of how emotions can be recognized and classified in the arena of spoken language. This algorithm relies on a highly curated dataset of emotive speech samples, each meticulously classified with an emotion descriptor, as its foundation.

This dataset serves as the foundation for a multi-step trip into the realm of SER, which includes data preparation, feature extraction, model training, and performance evaluation.The first phases of the programming entail importing this large dataset. Following that, it delves into the domain of data visualization, which is critical to understanding the emotional nuances expressed through speech. The algorithm reveals the specific acoustic elements that underlay each emotional state by displaying the waveforms and spectrograms of voice samples associated with various emotions. These visual cues serve as the foundation for deciphering the complicated interplay of auditory elements that communicate emotions and moods.

The feature extraction procedure is crucial in SER, and Mel- frequency cepstral coefficients (MFCCs) emerge as the preferred way for extracting relevant auditory information. MFCCs, a well-known approach, are computed from audio input, translating analog waveforms into a numerical representation that incorporates key acoustic properties. These extracted MFCCs are then used as input features for the emotion classification algorithm that follows.The emotion categorization model, which employs a deep learning architecture, specifically a Long Short-Term Memory (LSTM) neural network, is at the heart of the code. In this context, LSTM networks are suited for modeling sequential data and serve as the computational foundation for recognizing and discriminating emotional states in speech.

The model is painstakingly trained on the extracted MFCC features, and the emotion labels are encoded using one-hot encoding, allowing the network to learn and generalize from the dataset.The code meticulously monitors and assesses the model's performance as part of its development. To evaluate the model's accuracy in classifying emotions, loss and accuracy measures are used. This evaluation not only confirms the model's dependability, but also provides vital insights into its capacity to discern between various emotional states.The algorithm visualizes the probabilities associated with each emotion class to provide a comprehensive view of the model's predictions. This not only demonstrates the model's accuracy, but also its certainty in assigning emotions to speech samples.

II. RELATED WORK

The authors describe a novel method for improving overlapping voice recognition. They present an end-to-end architecture that integrates a multichannel neural speech extraction system with a deep acoustic model, doing away with the necessity for mel-filterbank (FBANK) extraction. They optimize the neural network simultaneously through recognition loss by combining these components into a single neural network. When compared to independently optimized systems, their technique leads in a significant 28% reduction in word error rate (WERR), displaying consistent performance across varied interference scenarios. They also simplify the model by replacing the FBANK extraction layer with a learnable feature projection, which results in better voice recognition. Furthermore, they objectively analyze speech quality in the reconstructed waveform from the enhancement network [1]. This paper presents the W-Net beamformer, a unique methodology that combines the strengths of deep neural network (DNN)-powered variants of generalized eigenvalue and minimum-variance distortionless response beamformers with DNN-based filter-estimation approaches. The proposed framework is divided into two parts: the first computes time-frequency references, and the second estimates beamforming filters based on these references. The results, collected from a variety of datasets with varying room and noise settings, show that the W-Net beamformer outperforms all assessment metrics. This implies that the suggested architecture allows for the efficient computing of beamforming filters, overcoming the

constraints of existing methods and presenting a promising improvement in the field of acoustic beamforming [2]. This letter introduces the Minimum Variance Independent Component Analysis (MVICA), a unique blind source separation (BSS) framework. MVICA strives for the highest possible output signal-to-interference ratio (SIR) while allowing greater flexibility in actual applications. Unlike standard BSS approaches, which rely on statistical independence, MVICA optimizes BSS algorithms based on the maximum SIR criterion, addressing circumstances where independence does not hold or is difficult to describe effectively. The letter also includes a deep neural network- supported MVICA implementation. Under various settings, experimental results show that MVICA surpasses state-of- the-art BSS algorithms in terms of SIR, signal-to-distortion ratio, and automatic speech recognition rate. This work marks an important step forward in BSS approaches [3]. This research describes a unique method for improving noise- robust automated speech recognition (ASR) by training a multi-channel beamformer alongside an acoustic model. It employs the complex ratio mask (CRM), which has been shown to be more successful than the ideal ratio mask (IRM), to estimate the covariance matrix of the beamformer. The CRM-based joint training framework is used to examine the Minimum Variance Distortion less Response (MVDR) and Generalized Eigenvalue (GEV) beamformers. In addition, the research presents a resilient mask pooling technique across several channels and uses a long short-term memory (LSTM) language model to re-score hypotheses, boosting overall performance. The approaches are tested using the CHiME-4 challenge dataset, and the CRM-based system achieves a 10% reduction in word error rate (WER) when compared to the IRM-based system [4]. This paper overcomes a limitation in previous work on neural network- based simultaneous multichannel augmentation and acoustic modeling. Existing techniques relied on predefined filters that were learnt during training, limiting adaptability to changing environments. To address this, the research proposes a mechanism called Neural Network Adaptive Beamforming (NAB). It use LSTM layers to anticipate time- domain beamforming filter coefficients for each input frame, allowing the filter to be dynamically adjusted. This modified filter is convolved with the input

signal and summed across channels before being processed by a waveform CLDNN acoustic model. When compared to a single-channel model, the NAB model produces a significant 12.7% relative improvement in Word Error Rate (WER) and performs similarly to a "factored" model that uses fixed spatial filters. It also cuts computational costs by 17.9% [5].

End-to-end training has proven effective in the context of neural beamformer-supported multi-channel Automatic Speech Recognition (ASR), but it confronts hurdles because to a lack of real-world multi-channel speech data. The purpose of this research is to investigate the use of single- channel data to improve multi-channel end-to-end ASR systems. Pre-training, data scheduling, and data simulation

are three ways for leveraging single-channel data. Experiments with the CHiME4 and AISHELL-4 datasets show that all three techniques improve training stability and ASR performance. In comparison to pre-training and data simulation, data scheduling is a simpler and less expensive strategy [6]. This study introduces Neural Architecture Search (NAS), which automates architecture design to improve end-to-end Automatic Speech Recognition (ASR). Recent gradient-based NAS methods, such as DARTS, SNAS, and Proxy less NAS, have improved NAS efficiency. The authors present ST-NAS, a NAS approach that uses Straight-Through (ST) gradients to optimize the loss through discrete variables, which was missing from Proxy less NAS. ST-NAS efficiently enables subgraph sampling, outperforming DARTS and SNAS. On the 80-hour WSJ and 300-hour Switchboard datasets, the study effectively applies ST-NAS to end-to-end ASR, displaying improved performance compared to human-designed architectures. ST-NAS is well-known for its architecture portability and low computational cost in terms of memory and time [7]. ESPRESSO is a Python-based open-source neural Automatic Speech Recognition (ASR) library based on FAIRSEQ. It supports distributed training over GPUs and nodes, as well as multiple ASR decoding algorithms. A fast, parallelized decoder for look-ahead word-based language model fusion is included. Without data augmentation, ESPRESSO beats other end-to-end ASR systems on datasets such as WSJ, LibriSpeech, and Switchboard. Furthermore, it decodes 4-11 times quicker than comparable systems like ESPNET, making it a cutting-edge solution for ASR jobs [8]. This paper introduces ESPnet, an open-source platform for end-to-end speech processing with a focus on automatic speech recognition (ASR). For deep learning, ESPnet makes use of popular dynamic neural network toolkits such as Chainer and PyTorch. It uses the Kaldi ASR toolkit approach for data processing, feature extraction, and recipe generation, providing a complete setting for speech recognition and other speech processing research. The paper describes the platform's design, significant characteristics that distinguish ESPnet from existing ASR toolkits, and experimental results obtained using major ASR benchmarks [9]. The self- attention transducer (SA-T) is introduced as a novel strategy for speech recognition in this paper, with the goal of addressing the parallelization difficulty faced by recurrent neural network transducers (RNN-T). SA-T substitutes RNNs with self-attention blocks, allowing for effective parallelization as well as the modeling of long-term relationships inside sequences. The research provides a path- aware regularization to improve alignment learning and performance. A chunk-flow method is also used for online decoding. Experiments on a Mandarin Chinese dataset (AISHELL-1) show a significant 21.3% reduction in character mistake rate when compared to the baseline RNN-
T. Furthermore, the chunk-flow approach of SA-T allows for online decoding with minimum performance impact [10].

This research investigates the use of transfer learning (TL) in end-to-end automated speech recognition (ASR) systems, with a focus on recurrent neural network transducers (RNN- T). In traditional hybrid ASR, TL is typically employed to transfer knowledge from a source language to a target language. The authors analyze four distinct TL approaches for RNN-T models in this work and discover that when compared to randomly initialized RNN-T models, these strategies result in a significant 10% to 17% relative reduction in word mistake rates. The study also investigates the efficacy of TL with varied quantities of training data, spanning from 50 to 1000 hours, highlighting the significance of TL in enhancing ASR performance for languages with low training data [11]. There is an increasing interest in building Recurrent Neural Network Transducer (RNN-T) models in the context of automated speech recognition (ASR). These models are trained without tight temporal alignment of transcripts and audio. Before producing ASR tokens, RNN-T models with unidirectional LSTM encoders frequently wait for longer input audio spans. This paper introduces Alignment Restricted RNN-T (Ar- RNN-T) models to overcome this issue. These models use audio and text alignment information to direct the loss computation, giving them more refined control over the trade-offs between token emission delays and Word Error Rate (WER). On datasets including Libri Speech and in- house data, comparisons with existing methods such as monotonic RNN-T are undertaken, illustrating the usefulness of the Ar-RNN-T loss in optimizing ASR performance [12]. The Recurrent Neural Network Transducer (RNNT)

is used in this study to train end-to-end voice recognition models. The RNNT is an all-neural, streaming architecture that learns both acoustic and language model components from transcribed acoustic input. The study investigates several model designs and shows how the model can benefit from additional text or pronunciation data. The model is made up of a 'encoder,' which is based on a connectionist temporal classification-based (CTC) acoustic model, and a 'decoder,' which is based on a recurrent neural network language model trained on text data. The complete neural network is trained with the RNN-T loss and generates the recognized transcript as a series of graphemes, allowing for end-to-end speech recognition [13]. The Recurrent Neural Network Transducer (RNNT) is used in this study to train end-to-end voice recognition models. RNNT is a streaming all-neural architecture that learns acoustic and language model components from transcribed auditory input. The study investigates several model designs and emphasizes the possible benefits of additional text or pronunciation data. The model consists of a 'encoder,' which is developed from a connectionist temporal classification-based (CTC) audio model, and a 'decoder,' which is affected by a recurrent neural network language model trained on text data. The RNN-T loss is used to train the whole neural network, which generates recognized transcripts as grapheme sequences, allowing for end-to-end speech recognition [14]. The research uses the wav2vec2 model for self-supervised learning (SSL) and investigates different pretraining and finetuning settings. The pretrained wav2vec2 models are fine-tuned utilizing various combinations of child and adult speech training data. The goal is to find the best data combination for fine-tuning the kid ASR model. The proposed model outperforms earlier techniques on datasets such as MyST, PFSTAR, and CMU KIDS in terms of Word Error Rates (WER). Notably, it outperforms the state-of-the- art wav2vec2 BASE 960 model built for adult speech recognition after only 10 hours of fine-tuning with child speech data [15].

This paper discusses the difficulty of customizing automated speech recognition (ASR) systems to better serve people with dysarthria, a speech disability. Dysarthric speech poses particular challenges, and collecting large training datasets is time-consuming for patients. To enhance ASR data, the study uses a Fast Speech 2-based multi-speaker text-to- speech (TTS) system to synthesis dysarthric speech. The study assesses the system's ability to generate dysarthric speech with little input (as few as 5 words) from an unknown speaker, and then uses it to train individualized speaker- dependent ASR models. While TTS quality could be improved, this approach has the potential to help in the future development of tailored acoustic models for new dysarthric speakers and domains [16]. The study addresses data scarcity through two degrees of augmentation. To begin, the original training data is improved by modifying prosody characteristics such as pitch and speaking rate. This enhancement boosts system performance. Second, the augmented data is utilized in conjunction with the original data to train a Text-to-Speech (TTS) system for generating synthetic data. The enlarged dataset is further enhanced by employing text-to-speech synthesis to generate children's utterances and diversifying the language model. When compared to the baseline system, the final speech recognition performance shows a significant relative improvement of 50.10% with acoustic diversity and 57.40% with language diversity-based augmentation [17]. This paper investigates the use of a hybrid of Convolutional Neural Networks (CNNs) and Transformer models for Automatic Speech Recognition (ASR). Transformers excel at capturing global interactions, whereas CNNs excel at managing local aspects. The suggested Conformer model integrates both, resulting in cutting-edge ASR performance. Conformer scores a Word Error Rate (WER) of 2.1%/4.3% on test other without a language model and 1.9%/3.9% using an external language model on the widely used Libri Speech benchmark. Even with a compact model with only 10 million parameters, a competitive performance of 2.7%/6.3% is achieved [18]. Spec Augment, a data augmentation technique for voice recognition, is introduced in the study. Spec Augment augments feature inputs directly, including feature warping, frequency channel masking, and time step masking. This approach is used for end-to-end speech recognition in Listen, Attend, and Spell networks. The results reveal that Spec Augment outperforms previous techniques and reaches state-of-the-art performance on tasks such as Libri Speech and Switchboard. Libri Speech obtains a WER of 6.8% on test-other without a language model and 5.8% with shallow language model fusion. It achieves 7.2%/14.6% WER without a language model and 6.8%/14.1% with shallow fusion on Switchboard, outperforming previous hybrid system performance [19]. This work addresses the problem of sequence transduction tasks like speech recognition and translation, where input and output sequences might be distorted in numerous ways. Traditional Recurrent Neural Networks (RNNs) are effective, but they necessitate a preset alignment of input and output sequences, which can be difficult. The study describes an end-to-end, probabilistic sequence transduction system based on RNNs that can convert any input sequence into any finite, discrete output sequence without the need for explicit alignment. The system's effectiveness is demonstrated by experimental findings on phoneme recognition using the TIMIT voice corpus [20].

A. ABBREVIATIONS AND ACRONYMS

1. ASR: Automatic Speech Recognition

2. WER: Word Error Rate

3. RNNT: Recurrent Neural Network Transducer

4. LSTM: Long Short-Term Memory

5. CNN: Convolutional Neural Network

III. METHODOLOGY

The mechanism for emotion identification from audio data is provided in this study. The first stage is to acquire an appropriate dataset containing audio recordings of emotional speech, which is then organized to provide structured access to audio files as well as their accompanying emotion labels. The data is then investigated and visualized using Python tools such as Pandas, NumPy, Matplotlib, Seaborn, and Librosa. To get insight into the features of the audio data, waveforms and spectrograms are formed. To verify the content and quality of the samples, audio playback is used. The extraction of Mel-frequency cepstral coefficients (MFCCs) from audio samples is a critical stage in the process. MFCCs are well-known for their capacity to capture significant information in speech and audio analysis, making them ideal for emotion recognition applications. The recovered MFCCs are processed, and emotion labels are encoded into one-hot encoded vectors to prepare the data for model training.

The heart of the methodology is the development of a neural network model for emotion recognition using the Keras framework. Layers in the model architecture comprise an LSTM (Long Short-Term Memory) layer for sequence

processing, followed by dense layers with activation functions. Dropout layers are intentionally inserted to prevent overfitting, and the final output layer includes softmax activation to categorize audio samples into one of seven emotion classes.

Model training then occurs, which entails compiling the model with appropriate loss functions and optimizers. The data is separated into training and validation sets, and the model is trained over a predetermined number of epochs and batch size. Throughout the training and validation process, loss and accuracy metrics are tracked. The model's performance is evaluated after training, and the probabilities predicted by the model for each emotion class are visualized for selected samples. Throughout the training, the patterns in training and validation loss and accuracy are examined. The performance of the model is evaluated using relevant evaluation metrics such as accuracy, F1-score, and confusion matrices.

The study concludes with a presentation and discussion of the findings, in which the model's performance is compared to existing methodologies, if relevant. Any difficulties, restrictions, or ideas for future work are thoroughly investigated. The study's conclusion highlights the important findings and contributions, providing insights into the research's implications in the context of emotion recognition and its prospective applications.

A. EQUATIONS

1. Discreate Fourier Transform:

$$xn=N1k=0\sum N-1Xke2\pi ikn/N$$

2. Mel Filterbank Operation:

$$mel(f)=2595log101+f700$$

3. Logarithmic Scaling:

y = loge(x) (blue)

B. ALGORITHMS

1. Feature Extraction Algorithms: The code's basic feature extraction approach is based on Mel- frequency cepstral coefficients (MFCCs). The Librosa library is used to extract MFCCs, which internally uses mathematical operations such as discrete Fourier transform (DFT), Mel-filterbanks, and logarithmic scaling to calculate the MFCC characteristics. These mathematical calculations are required for the MFCC feature extraction technique to work.

2. Machine Learning Algorithms: The code's primary machine learning method is a Long Short-Term Memory (LSTM) neural network. LSTM is a form of recurrent neural network (RNN) that works well with sequential data, such as audio. It is used to classify emotions based on the retrieved MFCC features. The code makes use of the Keras framework, which makes deep learning algorithms like LSTM easier to implement. The LSTM model is trained and optimized with the Adam optimizer with categorical cross-entropy as the loss function. The backpropagation technique is used in the training phase to iteratively update model weights and minimize the loss function.

3. Data Exploration and Visualization Algorithms: To generate charts for audio waveforms and spectrograms, the code leverages a variety of data visualization approaches. The Matplotlib and Seaborn packages were used to create these graphics. While not explicitly stated, these libraries generate plots internally using methods such as Fast Fourier Transform (FFT) for spectrograms and waveform charting algorithms for waveforms.

IV. IMPLEMENTATION

```
def extract_mfcc(filename):
    y, sr = librosa.load(filename, duration = 3, offset = 0.5)
    mfcc = np.mean(librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40).T, axis=0)
    return mfcc
```

```
extract_mfcc(df['Speech'][0])
```

```
array([-4.5629840e+02,  9.5333527e+01,  1.7716852e+01, -3.3765808e+01,
       -1.1817989e+01,  6.5464635e+00, -8.0160379e+00,  1.5736153e+00,
       -1.5325451e+01, -1.1126428e+00, -2.2056816e+01, -4.5272403e+00,
       -6.0576863e+00,  8.0293405e-01, -9.3291006e+00,  5.3059760e-02,
       -2.2767208e+00, -2.5546241e+00,  3.8283539e-01, -5.2236471e+00,
        3.8248415e+00,  2.9024797e+01,  2.4629841e+01,  3.5593071e+01,
        2.9523399e+01,  1.6402748e+01,  1.0630761e+00, -1.0897793e+00,
       -1.6727992e+00,  7.6605396e+00,  4.4564199e-01, -1.5233982e+00,
       -3.0100157e+00, -7.4013338e+00,  1.9788032e+00,  5.0565467e+00,
       -6.7794671e+00, -1.7020793e+00, -2.6077571e+00,  2.6367643e+00],
      dtype=float32)
```

```
X_mfcc = df['Speech'].apply(lambda x: extract_mfcc(x))
print("Features Extracted")
```

Model Architecture: The heart of our research is the architecture of the LSTM model. LSTM is a type of recurrent neural network (RNN) known for its ability to capture temporal dependencies in sequential data. The architecture consisted of a Sequential model with an LSTM layer as the primary component. The LSTM layer, configured with a specified number of units, served as the core element responsible for learning and representing sequential patterns in the audio data. The activation function employed in the LSTM layer was Rectified Linear Unit (ReLU), which introduced non-linearity and facilitated the model's ability to capture complex relationships in the audio features.

Additional layers were incorporated into the model, with Dense layers utilizing activation functions suitable for the problem at hand. For audio classification, the final layer employed the softmax activation function, allowing the model to output a probability distribution over the predefined emotion classes.

Model Compilation: Before training, the model was compiled by specifying essential parameters. The choice of loss function depended on the nature of the task, using mean squared error (MSE) for regression tasks and categorical cross-entropy for classification. The Adam optimizer was selected for optimization due to its efficiency in handling gradient updates. Training progress was monitored using evaluation metrics such as accuracy.

```
model = Sequential([
    LSTM(123, return_sequences=False, input_shape=(40,1)),
    Dense(64, activation='relu'),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dropout(0.2),
    Dense(7, activation='linear')
])

model.compile(loss='mean_squared_error', optimizer='adam', metrics=['mean_squared_error'])
```

Model Training: The training phase involved feeding the preprocessed audio data into the LSTM model. The model was trained on the training subset over a specified number of epochs, with a defined batch size. The objective was to minimize the chosen loss function while optimizing the model's ability to classify or predict audio emotions. The training process also included the validation of the model's performance using the testing subset.

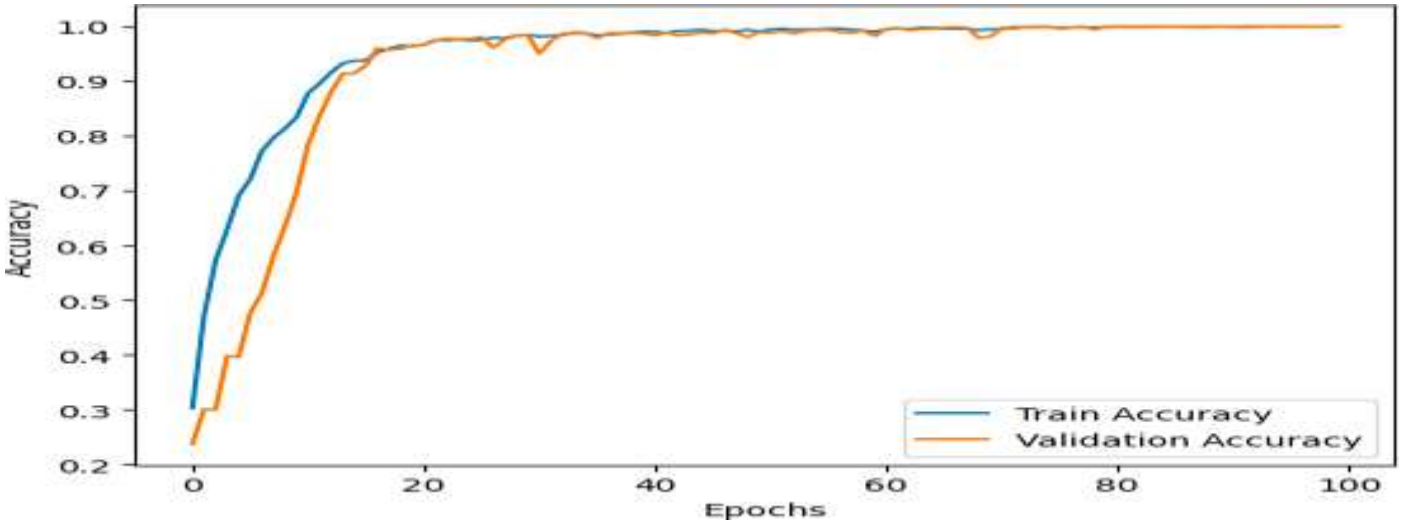
Model Evaluation: Following training, the model's performance was evaluated on the testing data. Evaluation metrics such as test loss and accuracy were employed to assess how well the model generalized to unseen audio samples. These metrics provided valuable insights into the model's ability to classify audio emotions.

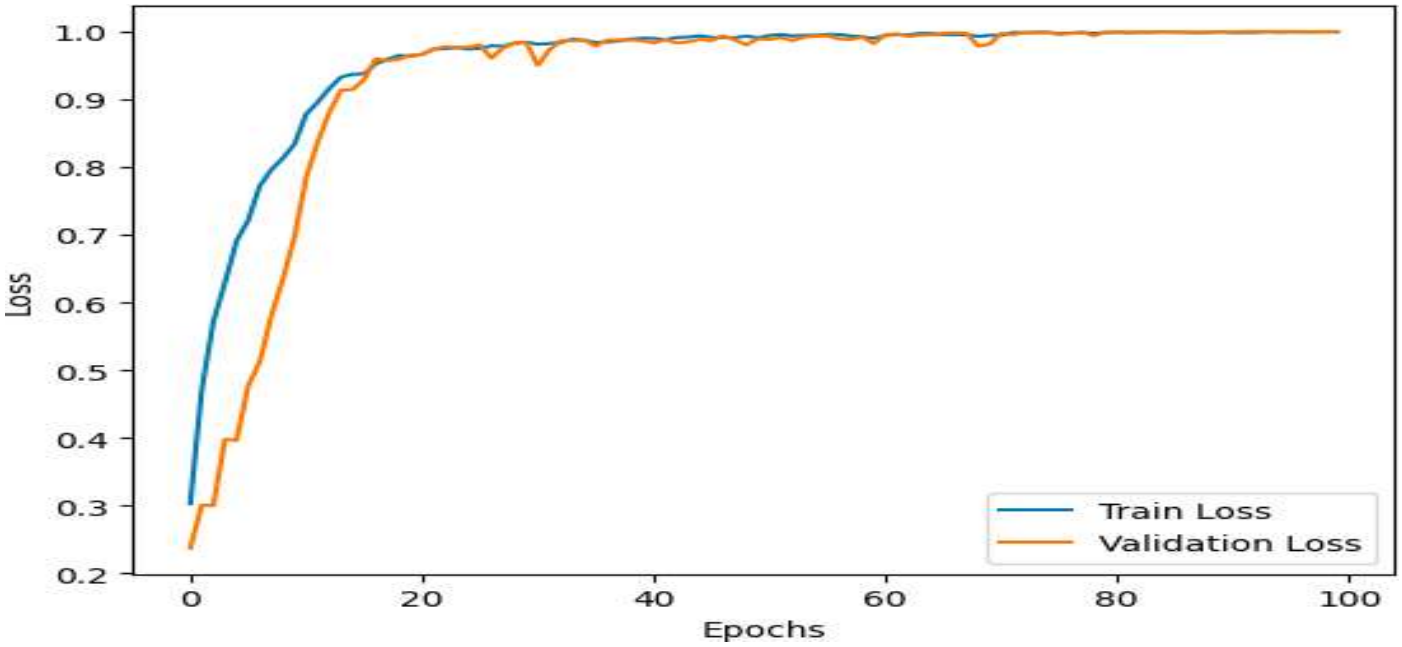
Outcome Visualization: To interpret the model's outputs, we visualized the predicted probabilities for each emotion class using bar plots. These visualizations allowed for a better understanding of how the model assigned probabilities to different emotion categories for a given audio sample.

In summary, our research employed a systematic approach that encompassed data preprocessing, LSTM model architecture, training, evaluation, and outcome visualization. These methodological steps enabled us to investigate the effectiveness of LSTM models for audio classification in the context of emotional speech analysis.

V. RESULT AND ANALYSIS

The result is plotted in a graph showing the accuracy percentages. The training accuracy being 99% and the validation accuracy is 100%.





VI. CONCLUSION

Finally, the offered code provides a full framework for emotion recognition from audio data by combining data preparation, feature extraction, and machine learning. The method starts by loading an audio dataset from Google Drive, then it explores the dataset's contents and distributes emotional labels. This visual examination provides researchers with vital insights about the structure and emotional content of the dataset. The feature extraction procedure, which is based on Mel-frequency cepstral coefficients (MFCCs), converts raw audio data into numerical representations that are machine learning friendly. These recovered features contain vital information about the spectral properties of audio signals, which is required for accurate emotion recognition. A Long Short-Term Memory (LSTM) neural network, a powerful machine learning method, lies at the heart of the code and is trained to categorize emotions based on the extracted MFCC data. For emotion recognition tasks, the model has been rigorously designed and tuned. While the code does not clearly reveal the mathematical equations involved, the code's libraries, such as Librosa and Keras, make use of the underlying methods for feature extraction and machine learning.The data visualization components of the code provide critical insights into the audio data, showing audio waveforms and spectrograms and allowing researchers to

assess the emotional content of the audio samples both qualitatively and statistically. This visual investigation ensures a thorough comprehension of the acoustic features of the dataset.

VII. FUTURE SCOPE

The future of voice emotion detection holds enormous promise for novel applications in fields ranging from mental health and education to human-computer interaction. These changes are expected to be accompanied by a greater emphasis on ethical issues, transparency, and the need to adapt to varied cultures and individual needs. The field is ripe for expansion and collaboration, with opportunities for scholars and practitioners to have a significant impact on technology and society.

REFERENCES

[1] Alex Graves "Sequence Transduction with Recurrent Neural Networks" 14 Nov 2012.

[2] Kanishka Rao, Hasim Sak, Rohit Prabhavalkar "EXPLORING ARCHITECTURES, DATA AND UNITS FOR STREAMING END-TO-END SPEECH RECOGNITION WITH RNN-TRANSDUCER" 12 Jan 2018.

[3] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unn, Nelson Enrique Yalta Soplin , Jahn Heymann , Matthew Wiesner , Nanxin Chen , Adithya Renduchintala , Tsubasa Ochiai " ESPnet: End-to- End Speech Processing Toolkit" 30 March 2018.

[4] Daniel S. Park∗, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, Quoc V. le "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition" 3 Dec 2019.

[5] Yiming Wang, Tongfei Chen, Hainan Xu, Shuoyang Ding, Hang Lv1, Yiwen Shao, Nanyun Peng, Lei Xie, Shinji Watanabe, Sanjeev Khudanpur "ESPRESSO: A FAST END-TO-END NEURAL SPEECH RECOGNITION TOOLKIT" 30 sep 2019.

[6] Zhengkun Tian, Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengqi Wen "Self-Attention Transducers for End-to-End Speech Recognition" 28 sep 2019.

[7] Bo Wu, Meng Yu, Lianwu Chen, Chao Weng, Dan Su and Dong Yu "OVERLAPPED SPEECH RECOGNITION FROM A JOINTLY LEARNED MULTI-CHANNEL NEURAL SPEECH EXTRACTION AND REPRESENTATION" 30 oct 2019.

[8] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, Ruoming Pang "Conformer: Convolution-augmented Transformer for Speech Recognition" 16 May 2020.

[9] Yuichiro Koyama, Bhiksha Raj "Exploring Optimal DNN Architecture for End-to-End Beamformers Based on Time-frequency References" 23 May 2020.

[10] Vikas Joshi, Rui Zhao, Rupesh R. Mehta, Kshitiz Kumar, Jinyu Li "Transfer Learning Approaches for Streaming End-to-End Speech Recognition System" 17 August 2020.

[11] Jay Mahadeokar, Yuan Shangguan, Duc Le, Gil Keren, Hang Su, Thong Le, Ching-Feng Yeh Christian Fuegen, Michael L. Seltzer "ALIGNMENT RESTRICTED STREAMING RECURRENT NEURAL NETWORK TRANSDUCER" 5 Nov 2020.

[12] Huahuan Zheng, Keyu An, Zhijian Ou "EFFICIENT NEURAL ARCHITECTURE SEARCH FOR END-TO- END SPEECH RECOGNITION VIA STRAIGHT- THROUGH GRADIENTS" 11 Nov 2020.

[13] Keyu An, Zhijian Ou "EXPLOITING SINGLE- CHANNEL SPEECH FOR MULTI-CHANNEL END-TO- END SPEECH RECOGNITION" 6 June 2021.

[14] Jianjun Gu, Longbiao Cheng, Dingding Yao, Junfeng Li, and Yonghong Yan "A Novel Blind Source Separation Framework Towards Maximum Signal-To-Interference Ratio" 8 march 2022.

[15] Rishabh Jain, Andrei Barcovschi, Mariam Yiwere , Dan Bigi, Peter Corcoran,Horia Cucu " A WAV2VEC2-BASED EXPERIMENTAL STUDY ON SELF-SUPERVISED LEARNING METHODS TO IMPROVE CHILD SPEECH RECOGNITION" 2017.

[16] Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez- Cabrero, Jesper N. Tegner "Whispering LLaMA: A Cross- Modal Generative Error Correction Framework for Speech Recognition" 16 october 2023.

[17] Enno Hermann and Mathew Magimai.-Doss " Few-shot dysarthric speech recognition with text-to-speech data augmentation" 20 August 2023.

[18] Bo Li, Tara N. Sainath, Ron J. Weiss, Kevin W. Wilson, Michiel Bacchiani "Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition" 2016.

[19] Yong Xu, Chao Weng, Like Hui, Jianming Liu, Meng Yu, Dan Su , Dong Yu " JOINT TRAINING OF COMPLEX

RATIO MASK BASED BEAMFORMER AND ACOUSTIC MODEL FOR NOISE ROBUST ASR"2018.

[20] Virender Kadyan, Hemant Kathania, Prajjval Govil, and Mikko Kurimo "Synthesis speech based data augmentation for low resource children ASR" 1 January 2021.