IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 7, July 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

Speech Emotion Recognition using Machine Learning

Nikhil T A, Mrs. Navya B J

Visvesvaraya Technological University, The National Institute of Engineering, Mysuru, India

Assistant Professor, Visvesvaraya Technological University, The National Institute of Engineering, Mysuru, India

ABSTRACT: One of the quickest and most natural ways for humans to communicate is through speech. Speech emotion recognition is the process of accurately anticipating a human's emotion from their speech. It improves the way people and computers communicate. Although it is tricky to annotate audio and difficult to forecast a person's sentiment because emotions are subjective, "Speech Emotion Recognition (SER)" makes this possible. Various researchers have created a variety of systems to extract the emotions from the speech stream. Speech qualities in particular are more helpful in identifying between various emotions, and if they are unclear, this is the cause of how challenging it is to identify an emotion from a speaker's speech. A variety of the datasets for speech emotions, its modelling, and types are accessible, and they aid in determining the style of speech. After feature extraction, the classification of speech emotions is a crucial component, so in this system proposal, we introduced Artificial Neural Networks (ANN model) that are utilised to distinguish emotions such as angry, disgust, Fear, happy, neutral, Sad and surprise.

I. INTRODUCTION

Emotions are vital to human communication, influencing interactions and decisions. Accurately recognizing emotions from speech is challenging due to individual variations in tone and speech patterns. Developing advanced systems to classify emotions from speech is crucial for various applications, such as improving human-computer interaction and understanding human psychology. Despite progress in machine learning, precise emotion detection remains a significant challenge.

II. OBJECTIVES

- Gaining a throughout knowledge of the connections that exists between the emotional states that speech signal acoustic qualities indicate and the signals themselves.
- Train and optimize machine learning models to accurately classify a diverse range of emotions based on speech features.
- Perform extensive evaluations and compare the resulting emotion detection models' performance with current techniques.
- Identify challenges and limitations in current emotion detection techniques and propose potential solutions or improvements for future research.

III. LITERATURE SURVEY

This literature survey covers various methods and approaches to gait analysis and recognition, highlighting advancements in the field across different applications.

1) A review on finding efficient approach to detect customer emotion analysis using deep learning analysis
2) Real Time Sign Language Recognition and Speech Generation
3) Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest
4) An i-vector GPLDA system for speech based emotion recognition
5) Reconstruction-error-based learning for continuous emotion recognition in speech

IV. EXISTING SYSTEM

Gamage et al. introduced a Gaussian-based approach for distinguishing emotional levels in speech, leveraging I-vectors to capture the distribution of MFCC features. Their study, utilizing the IEMOCAP corpus, demonstrates that the

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

GPLDA framework significantly surpasses the SVM framework in performance. Additionally, GPLDA shows greater robustness and less sensitivity to changes in I-vector dimensionality during system development.
In another study, Han et al. this paper introduced a novel approach to improve continuous emotion recognition in speech by employing a reconstruction-error-based learning framework that leverages memory-enhanced recurrent neural networks. Their approach employs two sequential RNNs: the first functions as an autoencoder to reconstruct the original input, and the second is used for emotion prediction. The reconstruction error generated by the first RNN is used as an additional feature, which, when combined with the original features, is fed into the second RNN to enchance the accuracy of emotion classification.

V. METHODOLOGY

Machine Learning
Machine Learning is a collection of computer calculations that can learn from cases and progress themselves without being expressly executed by a software engineer. Artificial intelligence incorporates machine learning, which uses information and statistical strategies to estimate an output that may be utilize to derive practical insights. The brain that does all the learning is machine learning. Machine learning is comparable to human learning. Experience teaches humans new things. Predicting becomes easier the more information we have. By similarity, our chances of victory are lower in an unknown situation than in a recognized one. Computers get the same training. The machine sees an case in arrange to create an exact expectation. When we give a comparable example to the machine it can figure out the result. Be that as it may, like a human, if its bolster a already inconspicuous case, the machine has troubles to anticipate.

The following points can be used to sum up the simple life of machine learning programs :
1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

The algorithm applies its learned conclusions to new data sets as soon as it becomes proficient at making the correct decisions.

Backpropagation to train multilayer architectures
Researchers have aimed to replace hand-engineered features with trainable multilayer networks since the early days of pattern recognition. Despite its simplicity, this approach wasn't widely understood until the mid-1980s, when it was discovered that multilayer architectures could be trained using stochastic gradient descent. The backpropagation procedure, a practical application of the chain rule for derivatives, allows for gradient computation through all layers by working backward from the output. This method was independently discovered by various groups in the 1970s and 1980s. The resurgence of interest in deep feedforward networks around 2006 was driven by the introduction of unsupervised learning procedures, allowing the creation of feature detectors without labeled data. This pre-training, followed by fine-tuning with standard backpropagation, proved effective in tasks such as handwritten digit recognition and pedestrian detection, particularly with limited labeled data. The advent of fast GPUs facilitated these developments, enabling significant advancements in speech recognition by 2009. The approach yielded record-breaking results and was quickly adopted for practical applications. Unsupervised pre-training helped prevent overfitting in smaller datasets, and convolutional neural networks (ConvNets) emerged as a superior architecture, widely adopted in the computer vision community.

Multilayer perceptron
A multilayer perceptron (MLP) is a name for an advanced feedforward artificial neural network, comprising of completely connected neurons with a nonlinear activation function, organized in at least three layers, notable for being able to recognize information that's not linearly distinguishable. The backpropagation calculation requires MLPs to utilize continuous activation functions such as ReLU.
Process for building MLP classifier has following steps:
- Declare MLP classifier by defining and the required parameters.
- To train classifier we are giving data to Neural Networks.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

• To predict the output, we use trained network.

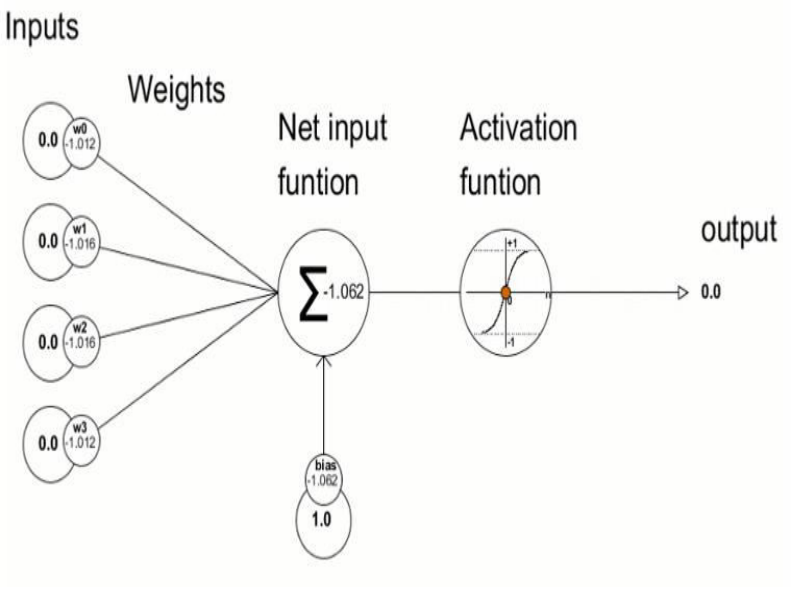

Fig 6.1 Multi Layer Percptron

VI. PROPOSED METHODOLOGY

The aim of computing is to enable efficient and natural human-computer interaction. One important objective is to make it possible for computers to comprehend the emotional states that people express so that tailored responses may be given. The majority of studies in the literature concentrate exclusively on recognizing emotions from short, isolated words, which prevents practical applications. We use artificial neural networks to implement voice emotion recognition in the suggested system (ANN model). The proposed system, which comprises of seven different emotion categories that is based on experiments using pre-recorded datasets.

The system requests training data, which includes weight training and expression labelling for that network. The input is an audio file. The audio is then subjected to intensity normalization. To prevent the impact of the presentation sequence of the samples from affecting the training performance, the ANN is trained using a normalized audio. The weight collections that are obtained as result of this training procedure that produce the best results when used with the learning data. The dataset retrieves the system with pitch and energy during testing, and based on the final network weights learned, it provides the identified emotion.

Advantages of proposed system

- ANN can provide data for processing in parallel, they can tackle multiple tasks at once.
- Resistance has been seen towards ANN so performance is thereby impacted by the loss of one or more neurons, or cells.
- ANN store information to generate results even when no data pairs are present.
- ANN are gradually being broken down, it will not cease to function abruptly, and that these networks will gradually degrade.
- We are able to train ANN's that these networks learn from past events and make decisions.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

Proposed Model Architecture

Proposed Model architecture

Data Collection

This block refers to collecting data which consists of 7 audio files with labeled emotions.
This block represents the raw audio data used to train the system. This data likely consists of speech recordings labelled with the corresponding emotions.

Data Preprocessing

Some audios are recorded at a variable rate, such as 44kHz or 22kHz. Librosa will run at 22KHz, and we will be able to observe data in a normalized pattern. Now, our motive is to extract some critical information while keeping our data in the form of independent (audio signal extracted characteristics) and dependent features (class labels). We'll employ MFCC to extract independent features from audio streams.

MFCCs

The MFCC summarizes the frequency distribution across the window size. So, it is possible to analyze both the frequency and time characteristics of the sound. This audio representation will allow us to identify features for classification. So, it will try to convert audio into some kind of features based on time and frequency characteristics that will help us to do classification Now, we have to extract features from all the audio files and prepare the data frame. So, we will create a function that takes the filename (file path where it is present). It loads the file using librosa, where we get 2 information. First, we'll calculate the MFCC of the audio data, and then we'll compute the mean of an array's

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

transposition to determine scaled features. To extract all of the features from each audio file, we must run a loop over each row in the data frame. We also use the TQDM Python package to monitor progress. Within the loop, we'll create a unique file path for each file before calling the method to extract MFCC features and attach them to a freshly generated data frame with appropriate labels.

Ann Model Creation

Split the dataset into train and test. 80% train data and 20% test data.

Now we will implement an ANN model using Keras sequential API. The number of classes is 7, which is our output shape (number of classes), and we will create ANN with 3 dense layers and architecture is explained below.

- The initial layer of the neural network comprises 100 neurons. Given 40 input features, the layer processes data with an input shape of 40. To introduce non-linearity, a ReLU activation function is applied. A dropout layer with a rate of 0.5 is incorporated to mitigate overfitting and enhance model generalization.
- The second layer has 200 neurons with activation function as Relu and the drop out at a rate of 0.5.
- The third layer again has 100 neurons with activation as Relu and the drop out at a rate of 0.5.

Compile the Model

To compile the model, loss function which is categorical cross-entropy are defined , accuracy metrics which is accuracy score, and an optimizer which is Adam.

Train the Model

The model will undergo training for 100 epochs, processing data in batches of 32 samples each. To monitor training progress and potentially save intermediate model states, a checkpoint callback is implemented. Upon completion, the trained model will be saved in the HDF5 format for future use.

Check the Test Accuracy

Finally, we will use the evaluate () method to assess the trained model on the test set and determine the model's correctness or Accuracy.

Tools and Technologies Required

The report covers the hardware and software requirements for the development of speech emotion recognition using machine learning are :

Hardware and Software Requirements

RAM	:	2 GB
Hard disk	:	100 GB
Process	:	32/64 Pentium

Software Requirements

IDE	:	FLASK
Language	:	Python.
Tool	:	Jupyter Notebook
Software	:	Anaconda
Front End	:	HTML, CSS, JavaScript
Libraries	:	Tensorflow, keras, numpy, pandas, skitlearn
Operating System	:	Windows/Android

VII. CONCLUSION

Speech Emotion Recognition (SER) constitutes a compelling domain within software engineering and artificial intelligence research. The proposed approach aligns closely with existing research endeavors in SER and offers a robust foundation for investigating the intricacies of emotion calculation from speech data. Subsequently, multilingual Emotion could be added to the suggested framework, expanding its usefulness in a variety of linguistic circumstances. The framework might also be improved to identify more complex emotional states, including tiny distinctions and complexities in emotional appearance. In addition to multilingualism, incorporating multimodal data—such as physiological signals and facial expressions—to enhance the precision and resilience of emotion recognition is an exciting new study direction. It may be possible to create more comprehensive emotion identification algorithms

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 7, July 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1207093|

through integration that are more like human emotional perception. These advancements have the potential to revolutionize relationships between humans and computers, enhance mental health applications, and promote a better comprehension of human emotions in diverse settings. Emotional data can be used to explore subtle nuances and recurring trends. This model was able to classify the emotion at 99.25% accuracy.

REFERENCES

[1] Kottursamy, Kottilingam. "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis." Journal of Trends in Computer Science and Smart Technology 3, no. 2 (2021): 95-113.
[2] Thakur, Amrita, Pujan Budhathoki, Sarmila Upret i, Shirish Shrestha, and Subarna Shakya. "Real Time Sign Language Recognition and Speech Generation." Journal of Innovative Image Processing 2, no. 2 (2020): 65-76.
[3] Kaur, Jasmeet , and Anil Kumar. "Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest." In Computer Networks and Inventive Communication Technologies, pp. 499-509. Springer, Singapore, 2021.
[4]Gamage, Kalani Wataraka, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah. "Ani-vector gplda system for speech based emotion recognition." In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 289-292. IEEE, 2015.
[5]Han, Jing, Zixing Zhang, Fabien Ringeval, and Björn Schuller. "Reconstruct ion-error-based learning for cont inuous emot ion recognition in speech." In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 2367-2371. IEEE, 2017.
[6] Akrami, N., F. Noroozi, and G. Anbarjafari. "Speech based emotion recognition and next reaction prediction." In 25th Signal Processing and Communications Applications Conference, Antalya, pp. 1-6. 2017.
[7] Rieger, S. A., Muraleedharan, R., & Ramachandran, R. P. (2014, September). Speech based emotion recognition using spectral feature extraction and an ensemble of kNN classifiers. In The 9th International Symposium on Chinese Spoken Language Processing (pp. 589-593). IEEE.
[8] Tabatabaei, Talieh S., and Sridhar Krishnan. "Towards robust speech based emotion recognition." 2010 IEEE International Conference on Systems, Man and Cybernetics. IEEE, 2010.
[9] Z. Li, "A study on emotional feature analysis and recognition in speech signal," Journal of China Institute of Communications, vol. 21, no. 10, pp. 18–24, 2000.
[10]T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.
[11] Vaijayanthi, S., and J. Arunnehru. " Synthesis Approach for Emotion Recognition from Cepstral and Pitch Coefficients Using Machine Learning." In International Conference on Communication, Computing and Electronics Systems, p. 515.
[12] Y. Kim, H. Lee, and E. M. Provost , "Deep learning for robust feature generation in audio-visual emotion recognition," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.
[13].Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391
[14] Chourasia, Mayank, Shriya Haral, Srushti Bhatkar, and Smita Kulkarni. "Emotion recognition from speech signal using deep learning." Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020 (2021): 471-481.
[15] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review", IEEE Access, vol. 2, no. 7, pp. 117327-117345, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details