International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/IJSRCSEIT

Performance Evaluation of Speech Emotion Recognition with Conventional Neural Network

Sana Fatema N. Ali[1], Prof. S. T. Khandare[2] , Prof. S. Y. Amdani[3]

[1]ME Scholar, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

[2]Associate Professor, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

[3]Associate Professor, Babasaheb Naik College of Engineering, Pusad, Maharashtra, India

ARTICLEINFO

Article History:

Accepted: 15 Oct 2023

Published: 10 Nov 2023

Publication Issue

Volume 9, Issue 6

November-December-2023

Page Number

37-47

ABSTRACT

The realm of speech emotion recognition presents a formidable challenge, offering valuable insights into the emotional states of speakers and facilitating enhanced human-machine interactions. However, in various scenarios, particularly those involving resource-constrained environments like embedded systems, the need arises to discern emotions in speech while grappling with limited computing and memory resources. While some prior research has shown promising recognition rates through transfer learning techniques utilising popular models such as Alex Net, a significant hindrance remains their substantial model size, rendering them impractical for execution on embedded systems. In response to this challenge, we present an innovative solution: a compact deep convolutional neural network architecture tailored to address the demands of resource-constrained environments.

Keywords- Speech emotion, Deep learning, CNN, MATLAB

I. INTRODUCTION

Emotions play a fundamental role in human communication, whether through spoken or written words. Effective communication hinges on our ability to grasp the emotional context conveyed by our interlocutors. While concealing emotions in written text is possible, psychological research has demonstrated the difficulty in masking feelings through physical and verbal expressions. Facial expressions and voice tones are highly reliable indicators of one's emotional state, a fact substantiated by numerous psychological and physiological studies.

Emotions have a profound impact on how we feel and act; they trigger and shape our facial expressions and vocal tones. For instance, fear prompts the release of adrenaline, preparing us to flee from danger, while joy and excitement are evident when conversing with loved ones, cuddling pets, or engaging in exhilarating activities like mountaineering or skydiving.

Recognizing emotional states in speech, a field known as Speech Emotion Recognition (SER), is not a novel concept in the domains of Artificial Intelligence and Machine Learning. Nevertheless, it remains a formidable challenge that warrants further attention. Convolutional neural networks (CNNs) represent a prominent type of artificial neural network and have played a pivotal role in advancing various domains, including face recognition, image segmentation, object recognition, handwriting recognition, and, notably, speech recognition. The name "CNN" is derived from the mathematical operation of convolution, which is frequently employed for efficient pattern recognition in data, particularly in the context of precise image classification.

Deep Learning can be likened to the human nervous system. Machine Vision Deep Learning models are designed to learn from a wealth of audio and image data, akin to a training dataset, to solve specific problems. These models mimic the human capacity for visual interpretation by simulating neural networks. Each node within the network operates much like a neuron in the human nervous system, and the deep learning models are essentially a subset of Artificial Neural Networks. Deep learning algorithms acquire an in-depth understanding of input data as it traverses through each layer of the network. The initial layers focus on detecting low-level features, such as edges, while subsequent layers integrate and refine these features to create a more abstract representation.

Deep learning models are adept at recognizing patterns in various digital forms of data, including images, audio, and sensor data. For prediction, we prepare the data through pre-training and construct training and testing sets where the outcomes are known. The objective is to identify the most optimal node that yields a satisfactory output. These nodes operate at different levels, progressively making predictions and selecting the most suitable predictions to yield the best-fit outcome. This process is akin to true machine intelligence, drawing inspiration from the intricate workings of the human nervous system.

II. LITERATURE REVIEW

Xinzhou Xu et al generalised the Spectral Regression model exploiting the joins of Extreme Learning Machines(ELMs) and Subspace Learning (SL) was expected to overlook the disadvantages of spectral regression-based Graph Embedding (GE) and ELM. Using the GSR model, in the execution of Speech Emotion Recognition (SER) we had to precisely represent these relations among data. These multiple embedded graphs were constructed for the same. Demonstration Speech Emotional Corpora determined the impact and feasibility of the techniques compared to prior methods that include ELM and Subspace Learning (SL) techniques. The system output can be improved by exploring embedded graphs at more precise levels. Only least-squares regression along with l2-norm minimization was considered in the regression stage [1]

Zhaocheng Huang et al use a heterogeneous token-used system to detect speech depression. Abrupt changes and acoustic areas are solely and collectively figured out in joins among different embedding methods. Contributions towards the detection of depression were used and probably various health problems that would affect vocal generation. Landmarks are used to pull out the information particular to an individual type of articulation at a time. This is a hybrid system. LWs and AWs hold various information. AW holds sections of the acoustic area into a single token per frame, and on the contemporary, the abrupt changes in speech articulation are shown by LWs. The hybrid join of the LWs and AWs permits the exploitation of various details, more specifically, articulatory dysfunction into conventional acoustic characteristics is also incorporated [2]

Peng Song offers a Transfer Linear Subspace Learning (TLSL) framework for cross-corpus recognition of speech. TLSL approaches, TULSL, and

TSLSL were taken in count. TLSL aims to extract robust characteristics representations over corpora into the trained estimated subspace. TLSL enhances the currently used transfer learning techniques which only focus on searching the most portable components of characteristics TLSL can reach even better results compared to the 6 baseline techniques with stats significance, and TSLSL gives better outcomes compared to TULSL, in fact, all the transfer learning is more accurate than usual learning techniques. TLSL significantly excels TLDA, TPCA, TNMF, and TCA, the excellent transfer learning techniques based on characteristics transformation. A big setback that these early transfer learning methods possess was that they concentrate on searching the portable components of characteristics that tend to ignore less informative sections. The less informative parts are also significant when it comes to transfer learning results experimented that TLSL is implemented for cross-corpus recognition of speech emotion [3]

Jun Deng et al focused on semi-supervised learning with automatic encoders of speech emotion recognition. Significantly work was on joining generative and discriminative training, by partially supervised learning algorithms designed to settings where non-labelled data was available. The process had been sequentially evaluated with 5 databases in different settings. The proposed technique enhances recognition performance by learning the prior knowledge from unlabelled data in conditions with a smaller number of liberated examples. These techniques can solve the problems in mismatched settings and incorporate the learnings from different domains into the classifiers, eventually resulting in outstanding performance. This shows that the model is having the capacity to make good use of the combination of labelled and unlabelled data for speech emotion recognition. The residual neural network displayed that intense architectures make the classifier beneficial to pull out complicated structure in image processing [4]

Ying Qin et al presented Cantonese-speaking PWA narrative speech which is the basis of a completely automated assessment system. Experiments on the text characteristics driven by the proposed data could detect the impairment of language in the aphasic speech. The AQ scores were significantly correlated with the text characteristics learned by the Siamese network. The improvised representation of ASR output was leveraged as the confusion network and the robustness of text characteristics were felicitated to it. There was an immediate requirement of improving the performance of ASR on aphasic speech for generation speech that has more robust characteristics. It was necessary for the databases of pathological speech and other languages to apply this proposed methodology. As seen clinically the most desirable one is automatic classification of aphasia variants along with this large-scale accumulation of data is needed substantially [5]

Z. T. Liu et.al proposed a framework of SER, firstly they built the initial feature set which was composed of speaker-independent features and speaker-dependent features by extracting features. Secondly, selecting features by using correlation analysis that consists of distance analysis, partial correlation analysis, bivariate correlation analysis, and the Fisher criterion, then the redundant speech emotional features were discarded. After that, the optimal feature subset was obtained. Finally, an extreme learning machine (ELM) decision tree was constructed for emotion recognition. The experiment results showed that the ELM is more suitable for the decision tree algorithm and the effectiveness of the feature selection based on correlation analysis and the Fisher criterion was fully verified [6]

Yelin Kim and Emily Mower Provost explore whether a subset of an utterance can be used for emotion inference and how the subset varies by classes of emotion and modalities. They propose a windowing method that identifies window configurations, window duration, and timing, for aggregating segment-level information for utterance-level emotion inference. The experimental results using the

IEMOCAP and MSPIMPROV datasets show that the identified temporal window configurations demonstrate consistent patterns across speakers, specific to different classes of emotion and modalities. They compare their proposed windowing method to a baseline method that randomly selects window configurations and a traditional all-mean method that uses the full information within an utterance. This method shows a significantly higher performance in emotion recognition while the method only uses 40–80% of information within each utterance. The identified windows also show consistency across speakers, demonstrating how multimodal cues reveal emotion over time. These patterns also align with psychological findings. But after all the achievement, the result is not consistent with this method [7]

A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen used a well-designed Convolutional Neural Network (CNN) architecture regarding the video based emotion recognition. They proposed the method named as HOLONET has three critical considerations in network design. (1) To reduce redundant filters and enhance the non-saturated non-linearity in the lower convolutional layers, they used modified Concatenated Rectified Linear Unit (CReLU) instead of ReLU. (2) To enjoy the accuracy gain from considerably increased network depth and maintain efficiency, they combine residual structure and CReLU to construct the middle layers. To broaden network width and introduce multi-scale feature extraction property, the topper layers are designed as a variant of the inception-residual structure. This method is more realistic than other methods here. It's focused on adaptability in real-time scenarios rather than accuracy and theoretical performance. Though its accuracy is also impressive, this method is applicable only in video based emotion recognition. Other types of data rather than video, this method can't produce results [8]

Y. Fan, X. Lu, D. Li, and Y. Liu. Proposed a method for video-based emotion recognition in the wild. They used CNN-LSTM and C3D networks to simultaneously model video appearances and motions. They found that the combination of the two kinds of networks can give impressive results, which demonstrated the effectiveness of the method. In their proposed method they used LSTM (Long Short Term Memory) - a special kind of RNN, C3D – A Direct Spatio-Temporal Model and Hybrid CNN-RNN and C3D Networks. This method gives great accuracy and performance is remarkable. But this method is much convoluted, time-consuming and less realistic. For this reason, efficiency is not that impressive [9]

Zixing Zhang, Fabien Ringeval, Eduardo Coutinho, Erik Marchi and Bjorn Schüller proposed some improvement in SSL technique to improve the low performance of a classifier that can deliver on challenging recognition tasks reduces the trust ability of the automatically labelled data and gave solutions regarding the noise accumulation problem - instances that are misclassified by the system are still used to train it in future iterations. They exploited the complementarity between audio-visual features to improve the performance of the classifier during the supervised phase. Then, they iteratively re-evaluated the automatically labelled instances to correct possibly mislabelled data and this enhances the overall confidence of the system's predictions. This technique gives the best possible performance using SSL technique where labelled data is scarce and/or expensive to obtain but still, there are various inherent limitations that limit its performance in practical applications. This technique has been tested on a specific database with a limited type and number of data. The algorithm which has been used is not capable of processing physiological data alongside other types of data [10]

Wei-Long Zheng and Bao-Liang Lu proposed EEG-based effective models without labelled target data using transfer learning techniques (TCA-based Subject Transfer) which is more accurate in terms of positive emotion recognition than other techniques used before. Their method achieved 85.01% accuracy. They used to transfer learning and their method includes three pillars, TCA-based Subject Transfer, KPCA-based

Subject Transfer and Transductive Parameter Transfer. For data pre-processing they used raw EEG signals processed with a band pass filter between 1 Hz and 75 Hz and for feature extraction, they employed differential entropy (DE) features. For evaluation, they adopted a leave-one subject-out cross-validation method. Their experimental results demonstrated that the transductive parameter transfer approach significantly outperforms the other approaches in terms of the accuracies, and a 19.58% increase in recognition accuracy has been achieved [11]

Thiang, et al. Presented speech recognition using Linear Predictive Coding (LPC) and Artificial Neural Network (ANN) for controlling movement of mobile robots. Input signals were sampled directly from the microphone and then the extraction was done by LPC and ANN [12]

Ms.Vimala.C and Dr.V.Radha proposed an independent isolated speech recognition system for Tamil language. Feature extraction, acoustic model, pronunciation dictionary and language model were implemented using HMM which produced 88% of accuracy in 2500 words [13]

Cini Kurian and Kannan Balakrishnan found development and evaluation of different acoustic models for Malayalam continuous speech recognition. In this paper HMM is used to compare and evaluate the Context Dependent (CD), Context Independent (CI) models and Context Dependent tied (CD tied) models from this CI model 21%. The database consists of 21 speakers including 10 males and 11 females [14]

Suma Swamy et al. Introduced an efficient speech recognition system which was experimented with Mel Frequency Cepstrum Coefficients (MFCC), Vector Quantization (VQ), HMM which recognize the speech by 98% accuracy. The database consists of five words spoken by 4 speakers ten time [15]

Annu Chaudhary et al. Proposed an automatic speech recognition system for isolated and connected words of Hindi language by using Hidden Markov Model Toolkit (HTK). Hindi words are used for dataset extracted by MFCC and the recognition system achieved 95% accuracy in isolated words and 90% in connected words [16]

Preeti Saini et al. Proposed Hindi automatic speech recognition using HTK. Isolated words are used to recognize the speech with 10 states in HMM topology which produced 96.61% [17]

Md. Akkas Ali et al. Presented automatic speech recognition techniques for Bangla words. Feature extraction was done by Linear Predictive Coding (LPC) and Gaussian Mixture Model (GMM). Totally 100 words were recorded 1000 times which gave 84% accuracy [18]

Maya Money Kumar, et al. Developed Malayalam word identification for speech recognition systems. The proposed work was done with syllable based segmentation using HMM on MFCC for feature extraction [19]

Jitendra Singh Pokhariya and Dr. Sanjay Mathur introduced Sanskrit speech recognition using HTK. MFCC and two state of HMM were used for extraction which produces 95.2% to 97.2% accuracy respectively [20]

Geeta Nijhawan et al. developed a real time speaker recognition system for Hindi words. Feature extraction done with MFCC using Quantization Linde, Buzo and Gray (VQLBG) algorithm. Voice Activity Detector (VAC) was proposed to remove the silence [21]

III. METHODOLOGY

In this project, we employ the Convolutional Neural Network (CNN) algorithm for the task of speech recognition. The process begins with the introduction of an audio clip, followed by feature extraction. Subsequently, we leverage the CNN to classify and recognize speech emotions. The speech emotion recognition application is executed using a convolutional neural network. Following is the architecture of the system:

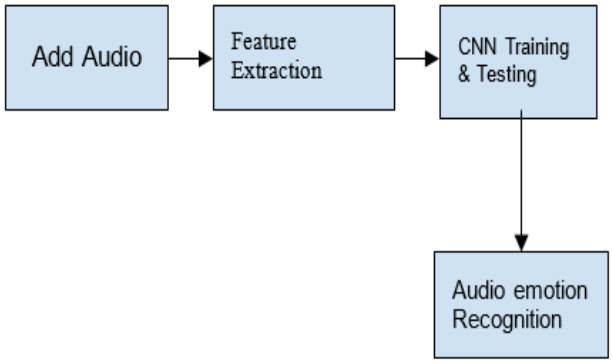


Fig 4: shows that System architecture Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a deep learning algorithm designed for processing and analysing visual data, particularly images and videos. It's a type of artificial neural network that's uniquely suited to tasks involving grid-like data, such as the pixel values in an image. CNNs have become a fundamental technology in computer vision, image recognition, and related fields. Here's how CNNs work:

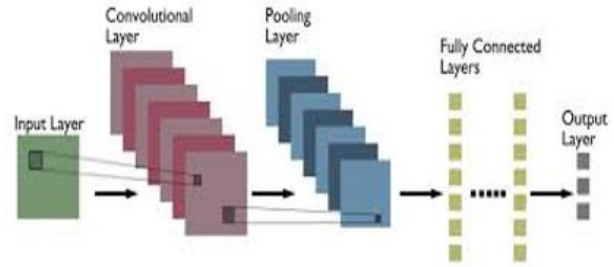


Fig 5: shows that the CNN model

1. Convolutional Layers: CNNs use convolutional layers to scan an input image or data grid. These layers consist of a set of learnable filters (also known as kernels) that are relatively small in dimension but extend across the full depth of the input. These filters move over the input data and perform element-wise multiplication, effectively learning to recognize patterns or features at various scales. The result is a feature map that highlights the presence of certain features in the input.

2. Pooling Layers: After convolution, CNNs often use pooling layers to down-sample the data and reduce its spatial dimensions. This helps in reducing the computational load and also makes the network more robust to variations in the input. Max-pooling and average-pooling are common pooling techniques used to select the maximum or average value in a local region of the feature map.

3. Fully Connected Layers: Once the convolution and pooling layers have extracted relevant features from the input, fully connected layers are used for making predictions or classifications. These layers connect all the neurons from the previous layer to the current layer, allowing the network to make high-level abstractions and make predictions based on the features extracted.

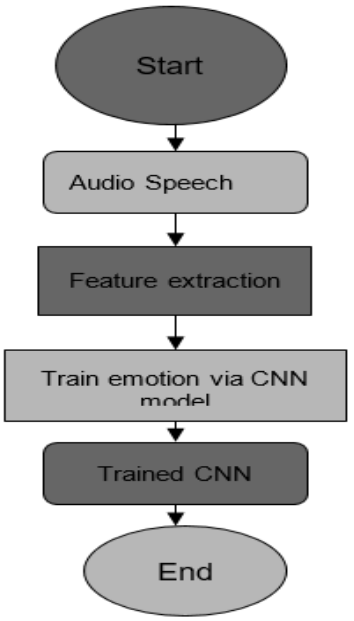
4. Activation Functions: Non-linear activation functions (e.g., ReLU - Rectified Linear Unit) are applied to the outputs of the convolutional and fully connected layers. These functions introduce non-linearity to the model, enabling it to learn complex patterns and make better predictions.

5. Backpropagation and Training: CNNs are trained using labelled data through a process known as backpropagation. During training, the network adjusts its parameters (the filter weights) to minimise the difference between its predictions and the actual labels in the training data. This process continues for multiple iterations (epochs) until the model converges to a state where it can make accurate predictions.

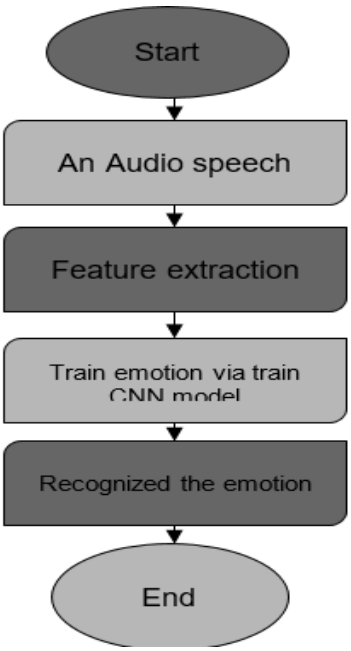
CNNs have proved highly effective in a wide range of tasks beyond image classification, including object detection, facial recognition, image generation, and more. They excel at learning hierarchical features from raw data and have played a pivotal role in the advancement of computer vision and deep learning.

FLOW CHART

Training flowchart



Testing flowchart



The model is initially trained using a dataset containing labelled expressions and associated training weights. An audio sample is used as input, and intensity normalisation is applied to standardise the audio. This normalised audio is employed to train the Convolutional Network, ensuring that the order of presentation of examples does not impact training performance. Through this training process, the model fine-tunes its weights, ultimately yielding optimal results with the training data.

During the testing phase, the system is provided with a dataset comprising pitch and energy information. Leveraging the final network weights established during training, the system determines and outputs the corresponding emotion associated with the input data.

Feature Extraction:

In the realm of speech recognition, the primary objective of the feature extraction process is to derive a concise sequence of feature vectors that encapsulate the essence of the input signal. Typically, this feature extraction occurs in three distinct stages. The initial stage, referred to as speech analysis or the acoustic front end, involves a spectral-temporal analysis of the signal. This analysis yields raw features that describe the power spectrum envelope of short speech segments.

In the second stage, an extended feature vector is constructed, comprising both static and dynamic features. Finally, in the last stage (which is not always included), these extended feature vectors are transformed into more compact and resilient vectors before being fed into the recognizer. While there isn't a consensus on the definitive characteristics of optimal feature sets, desirable properties typically include the ability to enable automatic systems to distinguish between similar-sounding speech sounds, facilitate the automatic generation of acoustic models for these sounds with minimal training data, and exhibit statistical consistency across various speakers and speaking environments.

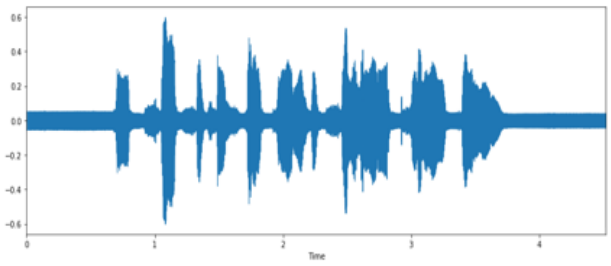


Fig 6 shows that feature Extraction

WORKING

The project titled "Speech Emotion Recognition with Conventional Neural Network" is a MATLAB-based initiative designed to recognize emotions conveyed in speech using Convolutional Neural Networks (CNN). The project's operation is characterised by a sequence of well-defined steps.

In this project, we commence by uploading an audio clip, serving as the input data. Subsequently, the feature extraction process is initiated to transform the raw audio data into a suitable format for analysis.
Following feature extraction, a crucial training phase is executed using CNN, which allows the system to learn and adapt to the distinctive patterns and nuances in speech associated with various emotions. Once the training is complete, the project transitions to the recognition phase. In this stage, when a user selects an audio clip from a set of eight options, the corresponding audio plays, and a corresponding line of text appears, articulating the emotion conveyed by the speaker during the speech. This system leverages the power of deep learning to analyse and categorise emotions in spoken language, offering an intuitive and practical tool for understanding the emotional content of speech.

In summary, this project employs a combination of audio input, feature extraction, CNN-based training, and emotion recognition to enable users to interactively explore and understand the emotions conveyed within speech samples, thereby enhancing our capability to interpret and respond to spoken language effectively.

STEP-BY-STEP WORKING:

Input Data: The project, titled "Speech Emotion Recognition with Convolutional Neural Network," commences by accepting an audio clip as input. This audio clip serves as the raw material for the subsequent analysis.

Feature Extraction: After the audio clip is uploaded, the project initiates the feature extraction process. This step transforms the raw audio data into a format that is more amenable to analysis. It's an essential preparatory step to extract relevant information from the audio.

CNN Training: Following feature extraction, a critical training phase is executed using a Convolutional Neural Network (CNN). The CNN is a deep learning model that specialises in recognizing patterns and nuances within the speech data that are associated with various emotions. Through training, the system learns to identify these emotional patterns and adapt its recognition capabilities accordingly.

Emotion Recognition Phase: Once the CNN training is successfully completed, the project enters the emotion recognition phase. In this stage, users can interact with the system by selecting an audio clip from a set of eight options. When a particular audio clip is chosen, the corresponding audio plays, and simultaneously, a line of text appears, explicitly articulating the emotion conveyed by the speaker during that specific speech segment.

Deep Learning Analysis: The system leverages the power of deep learning, specifically through CNN, to conduct a thorough analysis of the audio data. This deep learning approach allows the system to categorise and comprehend emotions within the spoken language,

offering an intuitive and practical tool for interpreting and grasping the emotional content of speech.

IV. SYSTEM REQUIREMENT

The system comprises mostly the software portion but has some hardware involved too. The Hardware that has been used was:

HARDWARE REQUIREMENT

1. CPU I3 processor
2. RAM 4GB
3. OS window 8
4. ROM 250 GB

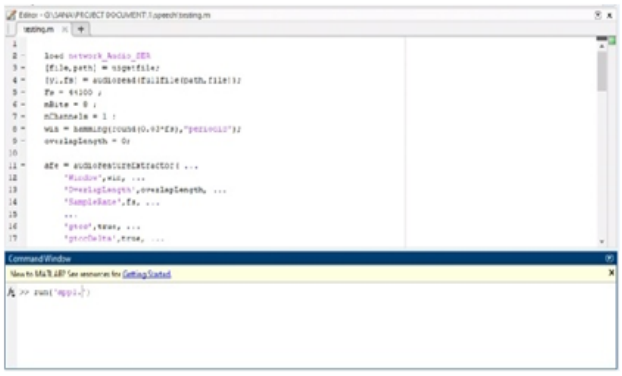
SOFTWARE REQUIREMENT

1. Matlab

IMPLEMENTATION

STEP 1: RUN THE CODE ON MATLAB

To execute your code, simply run the script by clicking the "Run" button in the MATLAB editor. This will initiate the process of loading and pre-processing data, designing and training the CNN, and evaluating the model's performance.

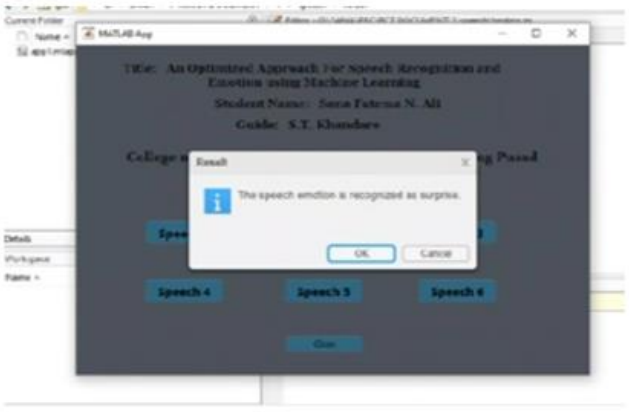


STEP 2: Data Pre-processing Load the speech dataset into MATLAB. Depending on your dataset, this might involve reading audio files, extracting features, and creating a labelled dataset.

Pre-process the audio data:

- Convert audio to spectrograms or MFCCs (Mel-frequency cepstral coefficients). These representations are commonly used for speech analysis.
- Normalise the data to have zero mean and unit variance.
- Split the data into training and testing sets.

STEP 3: SHOWING TEXT SHOWING EMOTION FROM AUDIO CLIP



V. CONCLUSION

After constructing various models, we got the better CNN model for the emotion distinction task. We reached 71%accuracy from the previously available model. Our model would've performed better with more data. Also, our model performed very well when distinguishing among a masculine and feminine voice.

VI. REFERENCES

[1]. X. Xu, J. Deng, E. Coutinho, C. Wu, and L. Zhao, "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," IEEE, vol. XX, no. XX, pp. 1–13, 2018.

[2]. Z. Huang, J. Epps, D. Joachim, and V. Sethu, "Natural Language Processing Methods for Acoustic and Landmark Event-based Features in Speech-based Depression Detection," IEEE J. Sel. Top. Signal Process. Vol. PP, no. c, p. 1, 2019.

[3]. P. S. Member, "Transfer Linear Subspace Learning for Cross-corpus Speech Emotion Recognition," vol. X, no. X, pp. 1–12, 2017.

[4]. J. Deng, X. Xu, Z. Zhang, and S. Member, "Semi-Supervised Auto encoders for Speech Emotion Recognition," vol. XX, no. XX, pp. 1–13, 2017.

[5]. Y. Qin, S. Member, T. Lee, A. Pak, and H. Kong, "Automatic Assessment of Speech Impairment in Cantonese-speaking People with Aphasia," IEEE J. Sel. Top. Signal Process. Vol. PP, no. c, p. 1, 2019.

[6]. M. D. Zeiler et al., "ON RECTIFIED LINEAR UNITS FOR SPEECH PROCESSING New York University, USA Google Inc., USA University of Toronto , Canada," pp. 3–7.

[7]. Yelin Kim and Emily Mower Provost, Data driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes; University of Michigan Electrical Engineering and Computer Science, Ann Arbor, Michigan, USA;2020.

[8]. A. Yao, D. Cai, P. Hu, S. Wang, L. Shan, and Y. Chen; HoloNet: towards robust emotion recognition in the wild, 2021

[9]. Y. Fan, X. Lu, D. Li, and Y. Liu.Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks. Proceedings of ICMI 2016 Proceedings of the 18th ACM International Conference on Multimodal Interaction, Pages 445-450,Tokyo, Japan — November 12 - 16, 2019.

[10]. Zixing Zhang, Fabien Ringeval, Fabien Ringeval, Eduardo Coutinho, Erik Marchi and Björn Schüller, Semi-Supervised Learning (SSL) technique.

[11]. Wei-Long Zheng1 and Bao-Liang Lu, Personalizing EEG-Based Affective Models with Transfer Learning, Centre for Brain-like Computing and Machine Intelligence, Department of Computer Science and Engineering, Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Brain Science and Technology Research Centre, Shanghai Jiao Tong University, Shanghai, China. 2018.

[12]. Thiang and Suryo Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robots", in Proceedings of International Conference on Information and Electronics Engineering (IPCSIT).

[13]. Ms.Vimala.C and Dr.V.Radha, "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", in Proceedings International Conference on Communication Technology and System Design 2020, Procedia Engineering 30 ISSN: 1877-7058, 13March 2020, pp.1097 – 1102.

[14]. Cini Kuriana, Kannan Balakrishnan, "Development & evaluation of different acoustic models for Malayalam continuous speech recognition", in Proceedings of International

Conference on Communication Technology and System Design 2020 Published by Elsevier Ltd, December 2020, pp.1081-1088

[15]. Suma Swamy, K.V Ramakrishnan, "An Efficient Speech Recognition System", Computer Science & Engineering: An International Journal (CSEIJ), Vol.3,No.4,DOI:10.5121/cseij.2019.3403 August 2021, pp.21-27

[16]. Annu Chaudhary, Mr. R.S. Chauhan, Mr. Gautam Gupta, "Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK)", in Proceedings of International Conference on Emerging Trends in Engineering and Technology, DOI: 03.AETS.2013.3.234, 22-24th February 2020, pp.244– 252.

[17]. Preeti Saini, Parneet Kaur, Mohit Dua, "Hindi Automatic Speech Recognition Using HTK", International Journal of Engineering Trends and Technology (IJETT)", Vol.4, Issue 6, ISSN: 2231-5381, June 2020, pp.2223-2229.

[18]. Akkas Ali, Manwar Hossain, Mohammad Nuruzzaman Bhuiyan, "Automatic Speech Recognition Technique for Bangla Words", International Journal of Advanced Science and Technology, Vol. 50, January, 2020, pp.51-60

[19]. Maya Money Kumar, Elizabeth Sherly, Win Sam Varghese, "Malayalam Word Identification for Speech Recognition System" An International Journal of Engineering Sciences, Special IssueiDravadian , Vol. 15 ISSN: 2229-6913 (Print), December 2021, pp. 22-26.

[20]. Jitendra Singh Pokhariya and Dr. Sanjay Mathur, "Sanskrit Speech Recognition using Hidden Markov Model Toolkit", International Journal of Engineering Research & Technology (IJERT),Vol.3, Issue 10, ISSN: 2278-0181, October-2020, pp.93-98

[21]. Geeta Nijhawan and Dr. M.K Soni, "Real Time Speaker Recognition System for Hindi Words", International Journal of Information Engineering and Electronic Business, Vol. 6, DOI: 10.5815/ijieeb.2019 .02.04, April 2020, pp. 35-40

Cite this article as :

Sana Fatema N. Ali, Prof. S. T. Khandare , Prof. S. Y. Amdani, "Performance Evaluation of Speech Emotion Recognition with Conventional Neural Network", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 6, pp.37-47, November-December-2023.