International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 6,Nov-Dec-2023, ISSN (Online): 2395-566X

Speech Emotion through Voice & Accent

Kundan Sai Kotta, Sai Nikhil Samineni, Asst. Prof.G. Kavitha
Dept. of AI&MLCBIT(A)
Hyderabad ,AP,India

Abstract- Detecting emotions through voice represents the next evolutionary leap in human-computer interaction, propelling us toward a more intuitive interface and enabling the development of superior recommendation systems.Voice, encompassing pitch, tone, and cadence, and accent, involving pronunciation patterns and linguistic nuances, play crucial roles in this context. Emotions, fundamental to human interaction, greatly influence communication and understanding. This research aims to investigate how variations in voice and accent contribute to expressing and interpreting emotions in speech. The study explores deep learning architectures and methodologies for this purpose, addressing associated challenges, limitations, and ethical considerations. Understanding the interplay of voice, accent, and emotions is pivotal for advancing technology in a beneficial manner.

Index Terms-voice, accent, emotion, intonation, deep learning.

I. INTRODUCTION

In general existing systems, be it advanced chatbots or language models, primarily emphasize the conversion of spoken words to text, neglecting the crucial layer of emotional nuances that significantly impact human communication. While the focus on accurate transcription is vital, understanding and interpreting the emotional content embedded within speech adds an invaluable dimension to human-computer interaction (HCI). Recognizing the emotional state conveyed through speech is pivotal for creating empathetic and contextually aware systems that can respond appropriately to users.Emotion plays a vital role in a human being's life. The requirement for human-to-computer communication had become unavoidable. To accomplish this, a computer would have to respond differently based on how it perceives the scenario in the present. To make human-computer interaction more natural, the computer must respond to human emotions in the same way as people in similar situations would. To achieve the goal, the computer can identify emotion through facial expressions or voices. Speech is a significant technique of recognizing emotions in HCI. SER has become one of the most important aspects of HCI[6].

To achieve this, we inherently face two major limitations with the available research. At present, researchers predominantly employ deep neural networks to train machine learning models for emotion classification. This approach offers notable advantages including rapid training, high classification accuracy, and enhanced capability compared to traditional machine learning methods [5]. Conversely, traditional machine learning techniques often encounter challenges such as local optimization issues and limited generalization ability[3]. In this research we try to establish a model that is capable of properly classifying emotions as set a standard of emotional state.When the speaker must repress emotions, some parts of internal sensation are buried and are not audible in speech. Therefore, computer-based systems are limited to what can be seen from the input of speech samples [6]. As a result of the lengthy dispute over the definition of "emotion" and the appropriate emotional classes, classifying emotional speech samples is a difficult task. To avoid that "fruitless discussion," Batlineretal. [7] favor the idea of emotion-related states [4]. However, among the systems that do acknowledge the importance of emotion detection, there remains a significant gap in addressing the diverse array of accents prevalent in the global population. Emotions are conveyed not only through words but also through variations in pitch, tone, and accent, which are unique to each individual and their cultural or regional background. Unfortunately, current emotion detection systems often fall short in effectively capturing these nuanced variations across accents.

But emotion detection from speech is quite difficult for many reasons: identifying the relevant emotion from a raw speech signal captured via a microphone can be affected by several factors such as gender, age, culture, health state, noise...The early Automatic Speech Recognition systems mainly focused on emotion recognition in several languages such as English[2]. Despite all the efforts, there has been little progress in determining which features were to select for improved performance [5].

Using a high-dimensional feature set that includes all sound parameters can aid in capturing all variances [6] but it can also lead to overfitting. A question could arise that whether a person expresses an emotion is largely dependent upon the person speaking, their culture, and the environment in which the person has been living. Majority of the study has concentrated on monolingual emotion classification, assuming that no cultural differences between speakers were present. Furthermore, application of large-scale acoustic parameters stood as a difficult task [7]. As a result of this, deep learning techniques are required for feature selection and low-latency SER. Through this study, we emphasize the significance of

not only recognizing emotions conveyed through words but also through variations in pitch, tone, and accent. Our approach seeks to bridge this gap, paving the way for more inclusive emotion detection systems. By employing deep learning techniques for feature selection and low-latency Speech Emotion Recognition (SER), we aim to develop a model that accurately captures emotional nuances across various linguistic and cultural backgrounds, ultimately advancing the field of Human-Computer Interaction (HCI).

II. LITERATURE SURVEY

- Dr. A. Arul Edwin Raj and Karan Kumar B have proposed a system where Mel-frequency Cepstral Coefficient (MFCC) feature is utilized to classify the data into different emotion groups. CNN is widely used for pattern recognition due to its many features like Mel Frequency Cepstral Coefficients (MFCC), a relatively simple structure, and fewer parameters for model training, making it ideal for SER. This technique effectively achieves a suitable compromise between the real-time process's performance precision and computing volume.As a machine learning model, the Speech Emotion Recognition (SER) system was developed.
- This research represents a new case study, aiming to construct and analyze an emotional speech corpus of the Algerian dialect. The objective is to propose a novel hybrid classification model designed to recognize emotions from Arabic speech. In pursuit of the research goals, a substantial annotated dataset comprising 1202 audio records was meticulously collected and constructed. These recordings were annotated with emotional labels such as happy, angry, neutral, or sad.Several experiments were conducted utilizing a variety of machine learning classification algorithms, in addition to deep convolutional and recurrent neural networks. It was observed that our proposed LSTM-CNN model surpassed all other classifiers and approaches, achieving an impressive accuracy of 93.34%. These results underscore the potential of LSTM networks in yielding compelling outcomes for speech emotion recognition. This is particularly significant in our case study, focusing on emotion detection from the Algerian dialect.
- Deep learning, a novel form of unsupervised methodology, employs artificial neural network models to analyze and process emotional information within the data. It utilizes data features in the deep learning process to identify and label the emotional content, thus enhancing understanding and analysis of the knowledge embedded within the dataset. This paper proposes a hybrid semantic text feature, integrating CNN and machine learning algorithms. The CNN algorithm is employed to analyze the data features, contributing to the development of a comprehensive and stable training model. Through human network training, the model is fine-tuned using CNN to label emotional content, and the data features generated during the deep learning process are utilized to configure the design parameters of the neural network model.
- Their custom feedforward-based deep learning model for speech emotion recognition demonstrated an impressive test accuracy of 93% and a training accuracy of 97.44%. Additionally, the test loss was recorded at 0.20, while the training loss stood at 0.081. Figure 6 in their study depicted the training and testing loss, along with the accuracy, illustrating a favorable performance during testing. The tools employed for simulation encompassed Python, Tensorflow, Numpy, Keras, and Google ColabPro.Their research primarily revolved around leveraging identifiers present in the datasets, including modality, vocal channel, emotional intensity, statement, repetition, and actor. These identifiers closely resembled stimulus characteristics. Notably, each expression was produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The overarching goal was to accurately label these emotions following a thorough assessment of the audio files. While the achieved results were commendable, they acknowledged potential areas for improvement. These areas included implementing noise filtering and conducting a more detailed analysis of the language used. They recognized that accurately determining the correct labels based solely on the provided identifiers could be challenging at times.
- In this experiment, the researchers applied speech emotion recognition to the IEMOCAP dataset. They utilized spectral time-frequency information as a feature extraction technology, incorporating various operations and filtering processes. For classification, deep learning technology was employed, utilizing a CNN model to effectively capture advanced features that preserve emotional characteristics in speech. Additionally, an LSTM model was used to maintain temporal information characteristics in speech.During the test stage, the Weighted Accuracy (WA) achieved a rate of 61%, and the Unweighted Accuracy (UA) stood at 56%. In their future studies, the researchers intend to shift their focus towards addressing neutral confusion and non-neutral confusion. They believe that resolving the neutral confusion problem could significantly enhance the recognition rate for emotion recognition.
- The application successfully achieved speech emotion recognition and underwent comprehensive testing for both functional and non-functional requirements, employing an ELM and RF classifier combination as its engine. The SER application met all 6 out of 6 defined requirements with 100% accuracy during the identification of 70 speech data instances.
- In the proposed method, CNN emerged as the superior and more reliable choice compared to previous architectures. The Speech Emotion Recognition (SER) model developed through this approach holds considerable significance in practical scenarios like police stations, car board systems, and call centers, where accurately interpreting emotions is vital. However, the

study encountered certain limitations, notably the challenge of handling large-scale acoustic parameters.

Additionally, the research primarily focused on monolingual emotion classification, assuming a homogeneity of cultural context among speakers.In acknowledging these limitations and aiming to enhance the model's versatility and accuracy, future research endeavors should address these concerns. Exploring various languages, such as Korean and Spanish, in the SER model is essential for broader applicability. Moreover, integrating LSTM and other deep learning models could potentially augment the precision and efficiency of the proposed method.

- The research team utilized Visual Studio to leverage obtained output for emotion detection using a trained model classifier, presenting the results accordingly. The study encompassed a comprehensive analysis of training and testing samples, along with the identified features. Accuracy and confusion matrix were derived by comparing predictions post-training showcasing an accuracy of 81.52% in the confusion matrix. This result indicates the model's effectiveness, with efforts underway to further improve prediction accuracy by extracting additional features.The research findings underscore the significance of this technique in the scientific and technical domain. Librosa was a key tool utilized to extract emotion recognition features, while Pyaudio facilitated audio recording. The Matplotlib module played a vital role in visualizing audio waves for future reference. Employing a classifier model, the team successfully categorized various emotions, setting a promising foundation for future advancements in this domain.

III. METHODOLOGY

The primary objective is to systematically collect, curate, and annotate new data tailored to suit identified algorithms for constructing the model. Extensive study of the data's features and their implications has informed the selection of specific models. Additionally, a thorough examination of the Evaluation Metrics outlined in various research papers has been conducted, with careful consideration given to integrating them for the robust evaluation of our model.

1.Data Collection

In the domain of data collection, our focal points centered on two critical aspects:

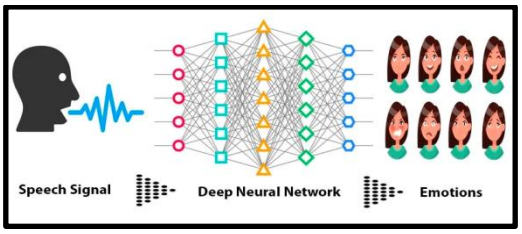
- Emotion-based Audio File Classification
- Incorporating Diverse Accent Variations within the Audio Dataset

All datasets consist of audio files. The audio files are processed by converting them into appropriate features, such as Mel-frequency cepstral coefficients (MFCCs), using relevant techniques. Necessary data pre processing, including augmentation and normalization, is performed using Keras. A comprehensive model architecture is proposed, incorporating Convolutional Neural Networks (CNN) for feature extraction from Mel-frequency cepstral coefficients (MFCCs) and Recurrent Neural Networks (RNN) for capturing sequential patterns. The model undergoes training on the curated dataset for a specified number of epochs. Evaluation involves testing the model with various audio inputs, including diverse accents, to assess its performance and achieve higher accuracy in voice emotion detection. The trained model is saved for future use, and a web application is developed using the Flask framework, enabling users to upload audio files for voice emotion analysis.

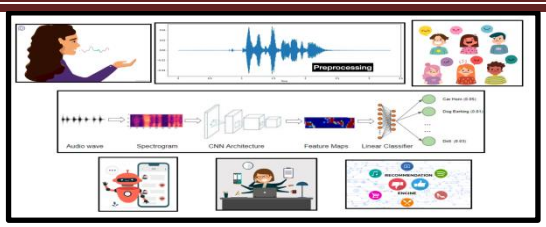
2.Speech Emotion Detection Process and Datasets

Analyzing speech emotion through voice and accent in real-time involves a structured process. The first step is data collection, gathering a diverse dataset of audio recordings encompassing various emotions and accents, utilizing platforms like Kaggle or recording custom data. Next, audio preprocessing is performed, extracting features like MFCC, pitch, and intensity analysis, and segmenting the audio into manageable units. Accent analysis is then implemented, utilizing linguistic analysis and pronunciation patterns to categorize accents. The subsequent step involves model selection and building, choosing appropriate deep learning models like CNNs, RNNs, or hybrids, and designing an architecture that integrates both emotion recognition and accent analysis aspects.

Afterward, the model is trained by splitting the dataset into training and testing sets, ensuring diverse representation, and using the chosen architecture and features. Integration into a real-time system follows, allowing live audio input or audio file processing. Real-time inference is then implemented, processing the audio input and providing predictions for both emotion and accent. Evaluation involves assessing model performance using metrics like accuracy, precision, recall, F1 score, and confusion matrix for both emotion and accent recognition, leading to fine-tuning based on results. Finally, integration into applications like customer service, virtual assistants, or mental health analysis is conducted, enabling a comprehensive system to recognize emotions and accents in real-time audio data, providing valuable insights and enhancing communication in various applications.



International Journal of Scientific Research & Engineering Trends
Volume 9, Issue 6,Nov-Dec-2023, ISSN (Online): 2395-566X



Datasets

Dataset Link	Locality
https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio	Well documented, clean dataset for UK dialect.
https://www.kaggle.com/code/lkergalipatak/speech-emotion-recognition-with-cnn/input?select=Crema	Augmented US dialect data.
https://www.kaggle.com/datasets/tapakah68/emotions-on-audio-dataset	GLOBAL
https://www.openu.ac.il/home/hassner	GLOBAL
https://roc-ng.github.io/XD-Violence	GLOBAL
http://en.arabicspeechcorpus.com/	ARABIC Phd Collection
https://gitlab.com/nicolasobin/att-hack/-/blob/master/README.md	FRENCH (Data)

3.Emotion detection using Deep Learning techniques

Deep learning has emerged as a powerful tool in recent years for predicting crime, leveraging various algorithms such as Convolutional Neural Networks (CNN), sentiment analysis, and deep neural networks. These algorithms are proficient in detecting patterns and anomalies in data, including text, images, audio, and social media, and can provide insights into potential criminal activities. When adapted for speech emotion detection through voice and accent analysis, these algorithms can prove instrumental.

- Customized CNN: CNN (Convolutional Neural Network) is a well-known algorithm used not only in image processing but also in speech emotion classification. In the context of analyzing audio data, the CNN model is customized to handle audio features extracted from speech signals. By assigning weights and biases to differentiate features relevant to various emotions, this CNN variant with 32, 64, and 128 filters effectively captures the emotional characteristics embedded in speech data. The model performs optimally on both training and testing datasets, highlighting its potential in speech emotion analysis.
- R-CNN: The concept of Region-based Convolutional Neural Network (R-CNN) is extended to speech emotion recognition, demonstrating its versatility beyond image analysis. By employing R-CNN architecture, features from different regions of the audio signal are extracted, enabling effective emotion recognition. The model, utilizing 32 filters and subsequent layers, can efficiently analyze audio data, providing valuable insights into the emotional content of speech. The integration of R-CNN in speech emotion recognition signifies a promising advancement in understanding and interpreting emotions through voice.
- VGGNET19 Adaptation: Adapting the VGGNET19 architecture, renowned for image classification, to the domain of speech emotion analysis showcases its applicability in various data types. In this context, VGGNET19 is tailored to handle audio features extracted from speech signals. The model, employing convolution layers and max-pooling, effectively processes the audio data, recognizing patterns that signify different emotions. This adaptation underscores the adaptability and effectiveness of deep learning architectures in analyzing emotional content in speech.
- ResNet50+LSTM Fusion: Combining Residual Network (ResNet) with Long Short-Term Memory (LSTM) networks presents a powerful approach for speech emotion classification. ResNet, known for its depth and accuracy, is utilized in feature extraction from audio data. The LSTM network, designed to analyze sequential data, effectively captures temporal dependencies in speech, enhancing emotion classification accuracy. The fusion of ResNet50 and LSTM offers a compelling solution for analyzing emotional nuances conveyed through speech signals.
- YOLOv5-inspired Approach : Inspired by the efficiency of YOLOv5 (You Only Look Once) in object detection, a similar approach is applied to real-time speech emotion recognition. The YOLOv5 architecture, designed for speed and accuracy, is adapted to efficiently process audio features extracted from speech signals. This innovative approach enables real-time analysis of emotional content in speech, showcasing the adaptability of cutting-edge techniques for audio-based applications.
- Simplified YOLO: The essence of the You Only Look Once (YOLO) algorithm, initially developed for object detection, is harnessed to efficiently detect and classify emotions in speech. This simplified YOLO model focuses on analyzing audio features encompassing emotional cues. By treating speech emotion detection as a regression problem, this approach combines efficiency with acceptable accuracy, showcasing its potential in real-time emotion analysis through speech.
- MobileNet: MobileNet, a lightweight deep learning architecture, is repurposed for speech emotion recognition. Its efficiency and computational speed make it an ideal choice for analyzing audio features extracted from speech. By processing audio data with precision and speed, MobileNet proves to be a valuable tool for real-time emotion

recognition in speech, paving the way for lightweight yet effective emotion analysis applications.

- Xception: Xception, known for its accuracy and efficiency in image recognition, is fine-tuned to handle audio features in speech emotion recognition. By leveraging its depth-wise separable convolutional network, Xception efficiently processes audio data, recognizing patterns indicative of various emotions. The adaptability of Xception in analyzing audio signals underscores its potential in enhancing speech emotion recognition accuracy.
- InceptionV3+LSTM: InceptionV3, recognized for its prowess in image recognition, is combined with LSTM to capture emotional patterns in sequential speech data. InceptionV3 extracts hierarchical features from audio signals, while LSTM analyzes sequential data to detect emotional nuances over time. This fusion provides a comprehensive approach to speech emotion recognition, offering insights into emotional variations within speech sequences.
- VGG16+LSTM: VGG16, a renowned deep CNN architecture, is paired with LSTM to track emotional dynamics within speech sequences. VGG16 effectively locates and tracks emotional cues, while LSTM identifies anomalies and patterns in speech sequences. This combined approach proves valuable in understanding how emotions evolve and manifest in speech over time, enhancing the accuracy of emotion classification in audio data.

IV.CONCLUSION

In striving to enhance human-computer interaction and accentuate the role of pitch and accent in communication, this research presents a method for real-time analysis of voice emotions. With a focus on achieving high training accuracy and minimizing loss during training, we propose a real-time system for emotion detection utilizing CNNs. This system adeptly discerns a range of emotions, even amidst diverse accents and varying pitch, enriching the user experience in human-computer interaction.

REFERENCES

1. Arul Edwin Raj, K. K. B, S. S and R. A, "Speech Emotion Recognition using Deep Learning," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 505-509, doi: 10.1109/ICIDCA56705.2023.10100056.
2. R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224.
3. W. Wang, G. Wen and Z. Zheng, "Design of Deep Learning Mixed Language Short Text Sentiment Classification System Based on CNN Algorithm," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5, doi: 10.1109/ICMNWC56175.2022.10031786.
4. D. Femi and S. Thylashri, "Human Voice Emotion Recognition Using Multilayer Perceptron," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-4, doi: 10.1109/ICSES55317.2022.9914336.
5. K. -Y. Huang, C. -H. Wu, Q. -B. Hong, M. -H. Su and Y. -H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5866-5870, doi: 10.1109/ICASSP.2019.8682283.
6. M. Saloumi et al., "Speech Emotion Recognition Using One-Dimensional Convolutional Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 212-216, doi: 10.1109/TSP59544.2023.10197766.
7. Ainurrochman, I. I. Febriansyah and U. L. Yuhana, "SER: Speech Emotion Recognition Application Based on Extreme Learning Machine," 2021 13th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2021, pp. 179-183, doi: 10.1109/ICTS52701.2021.9609016.
8. U. Mahesh YadavKonangi, V. R. Katreddy, S. K. Rasula, G. Marisa and T. Thakur, "Emotion Recognition through Speech: A Review," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 1150-1153, doi: 10.1109/ICAAIC53929.2022.9792710.
9. H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 795-799, doi: 10.1109/ICSIP52628.2021.9689043.
10. K. -Y. Huang, C. -H. Wu, Q. -B. Hong, M. -H. Su and Y. -H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5866-5870, doi: 10.1109/ICASSP.2019.8682283.