SPEECH EMOTION RECOGNITION

A. Ruthaya Trinishia

PG Department of Computer Science, St. Mary's College (Autonomous), Thoothukudi

Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India

ABSTRACT:

Emotion recognition from speech has emerged as a pivotal area of research with profoundimplications across various domains, ranging from human-computer interaction to mental health monitoring. The ability to decipher emotional cues embedded within speech not only enhances ourunderstanding of human communication but also holds the potential to revolutionize the way we interact with technology. In this context, the project titled "Speech Emotion Recognition" endeavors to develop an advanced system capable of accurately detecting and categorizingemotions expressed in audio samples. At its core, the project aims to bridge the gap between human emotion and machine intelligence by leveraging cutting-edge techniques from signal processing, deep learning, and natural language processing. By harnessing the power of data-driven approaches, the system seeks to extract meaningful features from audio signals and employ sophisticated algorithms to decode the underlying emotional states. Through a series of meticulously designed modules, including audio upload, playback, and emotion detection, users are provided with an intuitive platform to contribute to the dataset and interact with the system's functionalities seamlessly.

KEYWORDS:

Emotion recognition, Speech emotion, Deep learning, natural language processing, Detect speech emotion.

INTRODUCTION:

The significance of this project lies in its potential to revolutionize various facets of human- computer interaction. From virtual assistants capable of understanding and responding to users' emotional states to sentiment analysis tools facilitating deeper insights into customer feedback, theapplications are myriad. Moreover, in fields such as healthcare and education, the ability to accurately detect emotions from speech can aid in early diagnosis of mental health conditions, personalized learning experiences, and therapeutic interventions.

As technology continues to evolve, the need for empathetic and responsive systems becomes increasingly apparent. By delving into the realm of speech emotion recognition,

this project endeavors to lay the foundation for a future where machines not only understand what we say but also how we feel, paving the way for more empathetic and intuitive human-machine interactions.

The proposed system aims to develop a robust Speech Emotion Recognition (SER) model utilizing deep learning techniques. Leveraging advances in deep learning architectures, the system will process speech data to accurately detect and classify underlying emotional states. Initially, the system will preprocess raw audio data to extract relevant features, such as Mel-Frequency Cepstral Coefficients (MFCCs) or spectrograms, which capture key characteristics of speech signals. These features will then serve as input to a deep learning model, possibly a Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), or a combination thereof, designed to effectively learn intricate patterns and temporal dependencies within the data. Training the model will involve utilizing labeled datasets encompassing various emotional expressions across different speakers and contexts.

Additionally, techniques like data augmentation and transfer learning may be employed to enhance model generalization and performance. The system will undergo rigorous evaluation through metrics such as accuracy, precision, recall, and F1-score, ensuring its efficacy in real- world scenarios. As technology continues to evolve, the need for empathetic and responsive systems becomes increasingly apparent. By delving into the realm of speech emotion recognition, this project endeavors to lay the foundation for a future where machines not only understand what we say but also how we feel, paving the way for more empathetic and intuitive human-machine interactions.

MATERIALS AND METHODS:

Dataset Description:

The project entitled "Speech Emotion Recognition" utilized the following datasets:

- CREMA-D
- RAVDESS
- SAVEE
- TESS

1. Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D is a data set of 7,442 original clips from 91 actors. These clips were from 48 male and43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors

spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified).

2. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

Speech audio-only files (16bit, 48kHz .wav) from the RAVDESS. Full dataset of speech and song, audio and video (24.8 GB) available from Zenodo. Construction and perceptual validation of the RAVDESS is described in our Open Access paper in PLoS ONE. This portion of the RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professionalactors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgustexpressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

3. Surrey Audio-Visual Expressed Emotion (SAVEE)

The SAVEE database was recorded from four native English male speakers (identified as DC, JE,JK, KL), postgraduate students and researchers at the University of Surrey aged from 27 to 31 years. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise. A neutral category is also added to provide recordings of 7 emotion categories. The text material consisted of 15 TIMIT sentences per emotion: 3 common, 2emotion-specific and 10 generic sentences that were different for each emotion and phonetically-balanced. The 3 common and 2 × 6 = 12 emotion-specific sentences were recorded as neutral to give 30 neutral sentences. This resulted in a total of 120 utterances per speaker.

4. Toronto emotional speech set (TESS)

There are a set of 200 target words were spoken in the carrier phrase "Say the word _'by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800data points (audio files) in total. The dataset is organized such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

Project Description:

The project entitled "Speech Emotion Recognition" has the following modules:

- Upload Audio
- Play Audio
- Detect Audio

1. Upload Audio

This module enables users to upload audio files easily. Users can submit their recorded speech or conversation samples to the system for emotion analysis. The module ensures a user- friendly experience, allowing individuals to seamlessly contribute to the dataset used for training the speech emotion recognition model.

2. Play Audio

The Play Audio module allows users to listen to the uploaded audio samples directly within the application interface. This feature facilitates a quick review of recorded content and helps users confirm the accuracy of the uploaded data. It enhances user engagement by providing a convenient way to interact with the processed audio files.

3. Detect Audio

The Detect Audio module is the core of the speech emotion recognition system. Using deep learning algorithms, this module analyzes the uploaded audio samples to identify and categorize the emotional content. To detect audio the following steps are performed,

i. Feature extraction:

- **Zero Crossing Rate (ZCR):** Calculate the rate at which the audio signal changes its sign. This is done by counting the number of times the signal crosses zero within a given frame. ZCR provides insights into the speech signal's abrupt changes and can be indicative of certain emotional expressions.
- **Chroma Shift:** Utilize chroma features to represent the distribution of energy across different pitch classes. Chroma Shift involves shifting the chroma feature vectors to capture temporal variations in pitch content. This helps in capturing musical characteristics and emotional nuances related to tonal changes in speech.
- **Mel-Frequency Cepstral Coefficients (MFCC):** Apply the Fourier Transform to convert the audio signal into the frequency domain. Then, the Mel filter banks are used to extract the MFCCs, representing the spectral characteristics of the signal. MFCCs are widely used in speech processing for their effectiveness in capturing relevant information related to speech

production.

- **Root Mean Square (RMS) Value:** Compute the RMS value of the audio signal within each frame. RMS provides information about the overall energy level of thesignal. In speech emotion recognition, variations in energy levels can be associatedwith the intensity of emotional expression.
- **Mel Spectrogram:** Generate a Mel spectrogram by applying the Mel filter banks to the power spectrum of the audio signal. The Mel spectrogram represents the distribution of energy across different frequency bands over time, providing a detailed view of the spectral content. It helps in capturing both the temporal and frequency characteristics of the speech signal.

ii. Stretching and pitching:

Stretching and pitching are crucial steps in the speech emotion recognition process. During stretching, the temporal characteristics of the audio signal are adjusted, altering the duration of speech without changing its pitch. This step helps standardize the time domain features, ensuring consistency in the analysis. On the other hand, pitching involves modifying the frequency of the audio signal while maintaining its duration. This adjustment in pitch allows the system to focus on variations in vocal tone, which can be indicative of different emotions. Together, stretching and pitching contribute to the extraction of relevant features from the speech signal, enhancing the deep learning model's ability to discern subtle nuances in emotional expression and improving the overall accuracy of emotion recognition.

iii. CNN:

The audio data undergoes a series of transformations to extract meaningful features for emotion analysis. Initially, the raw audio signals are converted into spectrograms or Mel-frequency cepstral coefficients (MFCCs), which represent the frequency content over time. These spectrograms or MFCCs serve as input to the CNN. The network comprises convolutional layers that apply filters to detect local patterns and hierarchical features within the spectrogram, capturing both low and high-level representations. Pooling layers are then employed to down-sample the spatial dimensions, reducing computational complexity while retaining essential features. The processed information is flattened and fed into fully connected layers for classification. The CNN learns to recognize patterns associated with different emotions during the training phase, adjusting its parameters to optimize

performance. Once trained, the CNN can efficiently classify emotional states in unseen audio data, providing a powerful tool for Speech Emotion Recognition.

iv. Speech emotion recognition:

Angry, calm, disgust, fear, happy, neutral, sad, and surprise are the labels assigned to audio classesthat are identified in the Detect Audio module.

RESULTS AND DISCUSSION:

The "Speech Emotion Recognition" project exhibits promising results in accurately detecting and categorizing emotional content within audio samples. Through the integration of various modules, the system offers a comprehensive framework for users to upload, review, and analyze audio data seamlessly.

The "Upload Audio" module provides users with a straightforward interface to contribute audio samples, enhancing the dataset used for training the emotion recognition model. This feature promotes user engagement and ensures a diverse range of data for robust model training. In the subsequent "Play Audio" module, users can conveniently listen to uploaded samples directly within the application interface. This functionality facilitates quick review and validation of the recorded content, enabling users to confirm the accuracy of their contributions. Moreover, it promotes a user-friendly experience by providing easy access to processed audio files. The core of the system lies in the "Detect Audio" module, where deep learning algorithms are employed to analyze uploaded samples and categorize emotional expressions. Through a series of steps including feature extraction, stretching, pitching, and CNN-based analysis, the system effectively captures and interprets subtle nuances in speech signals associated with various emotions.

Feature extraction techniques such as Zero Crossing Rate (ZCR), Mel-Frequency Cepstral Coefficients (MFCC), and Mel Spectrogram enable the system to extract relevant information from the audio signal, representing both temporal and spectral characteristics. These features provide valuable insights into the speech signal's properties, aiding in emotion classification. The CNN architecture plays a pivotal role in learning hierarchical representations from the extracted features, allowing the model to discern patterns associated with different emotions. By leveraging convolutional and pooling layers, the CNN effectively captures local patterns and spatial information within spectrograms or MFCCs, enabling accurate classification.

The system successfully categorizes audio samples into eight emotional classes: Angry, Calm, Disgust, Fear, Happy, Neutral, Sad, and Surprise. This comprehensive range of emotion labels ensures that the model can accurately capture a wide spectrum of

emotional expressions, enhancing its utility in real-world applications.

Overall, the "Speech Emotion Recognition" project demonstrates significant potential in accurately identifying and categorizing emotional content within audio samples. By leveraging deep learning techniques and comprehensive feature extraction methods, the system provides a robust framework for emotion analysis, with implications for various applications including sentiment analysis, virtual assistants, and human-computer interaction. Further refinement and validation of the model with larger and more diverse datasets could potentially enhance its performance and generalizability in real-world scenarios.

CONCLUSION:

The "Speech Emotion Recognition" project presents a comprehensive framework for analyzing and categorizing emotional content in audio samples. Through a series of meticulously designed modules, including Upload Audio, Play Audio, and Detect Audio, users are provided witha seamless and intuitive interface for contributing to the dataset and interacting with the system's functionalities. The core module, Detect Audio, leverages advanced deep learning techniques such as feature extraction, stretching, pitching, and Convolutional Neural Networks (CNNs) to discern subtle emotional nuances embedded within speech signals. By extracting relevant features and employing a CNN architecture trained on diverse emotional states, the system demonstrates robust performance in accurately classifying emotions including angry, calm, disgust, fear, happy, neutral,sad, and surprise. This project not only showcases the potential of deep learning in understanding human emotions but also underscores its practical applications in various domains such as mentalhealth monitoring, human-computer interaction, and sentiment analysis in customer feedback. Moving forward, continued research and development in this field hold the promise of further enhancing the accuracy and versatility of speech emotion recognition systems, ultimately enriching human-machine interactions and fostering more empathetic and responsive technologies.

REFERENCES:

1. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M.A. and Cleder, C., 2019. Automatic speech emotion recognition using Deep Learning.
2. Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H. and Alhussain, T., 2019. Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, pp.117327-117345.

3. Anusha, R., Subhashini, P., Jyothi, D., Harshitha, P., Sushma, J. and Mukesh, N., 2021, June. Speech emotion recognition using Deep Learning. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1608-1612). IEEE.
4. Aouani, H. and Ayed, Y.B., 2020. Speech emotion recognition with deep learning. Procedia Computer Science, 176, pp.251-260.
5. Harár, P., Burget, R. and Dutta, M.K., 2017, February. Speech emotion recognition with deep learning. In 2017 4th International conference on signal processing and integrated networks (SPIN) (pp. 137-140). IEEE.
6. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Weninger, F. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In Proceedings of INTERSPEECH.
7. Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462).
8. Scherer, K. R., & Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal?. Emotion, 7(1), 113.
9. Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large- scale hierarchical image database. In 2009 IEEE conference on computer vision and patternrecognition (pp. 248-255).
10. Huang, Y., Chiu, C. C., & Wu, S. (2014). Deep learning for speech emotion recognition: An overview of recent developments. IEEE Access, 2, 530-541.
11. Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1175-1191.
12. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan,
    S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Languageresources and evaluation, 42(4), 335.
13. Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649).

14. Ringeval, F., Schuller, B., Valstar, M., Cowie, R., & Pantic, M. (2018). AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In Proceedings of the 2018 on audio/visual emotion challenge and workshop (pp. 3-9).
15. Liu, J., & Deng, L. (2018). Deep learning for speaker recognition: A comprehensive review. Speech Communication, 99, 207-219.