**Project Title -** Speech Emotion Recognition using Machine Learning.

**Abstract**.

Speech Emotion Recognition (SER) is an emerging field in human-computer interaction, aiming to identify and classify human emotions from speech signals. This paper proposes a novel SER method leveraging advanced feature extraction and machine learning techniques, demonstrating great performance on RAVDESS dataset using machine learning classifiers like MLP, SVM.

**Introduction**

Speech Emotion Recognition (SER) is a rapidly evolving field within human-computer interaction, aiming to identify and classify human emotions from speech signals. This educational paper delves into the complexities of SER, exploring advanced feature extraction techniques and machine learning models to enhance emotion detection accuracy. The paper provides a comprehensive overview of SER, its challenges, and potential solutions, making it a valuable resource for researchers and practitioners in the field.

**Dataset**

The RAVDESS dataset, used in this educational work, is a widely recognized resource in the field of Speech Emotion Recognition (SER). It contains a diverse collection of audio-visual clips featuring professional actors expressing various emotions through speech. This dataset provides a robust foundation for training and evaluating SER models due to its high-quality recordings and variety of emotional expressions.

**Methodology**

The methodology employed in this educational paper for Speech Emotion Recognition (SER) involves a two-step process. First, features are extracted from the audio file using Librosa, a Python library for music and audio analysis. This step allows for the conversion of complex audio data into a format that can be used for analysis. Following feature extraction, the data is classified using two machine learning models: Multilayer Perceptron (MLP) and Support Vector Machine (SVM). These models are trained to identify and classify emotions based on the extracted features, providing the foundation for emotion recognition. This methodology offers a comprehensive approach to SER, combining advanced feature extraction with robust classification techniques.

**Results**

In our Speech Emotion Recognition (SER) study, both Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) were employed. While SVM showed promising results, MLP outperformed it, proving to be more effective in emotion classification tasks.

**Conclusion**

In conclusion, our work on Speech Emotion Recognition (SER) has demonstrated the potential of machine learning models in emotion classification. The superior performance of the MLP model over SVM underscores the importance of selecting the right model for SER tasks.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Ref No** | **Year of Publication** | **Proposed work** | **Data set Used** | **Ml Algo** | **Advantage** | **Disadvantage** | **Accuracy** | **Brief Description** |
| 1 | [2] | 2022 | Individual Standardization Network (ISNet): Enhancing Speech Emotion Recognition by Alleviating Interindividual Emotion Confusion |  |  |  |  | 60.25% | The authors propose an Incomplete Sparse LSR (ISLSR) model to establish a linear relationship between speech features and corresponding emotion labels |
| 2 | [3] | 2020 | “Speaker-Specific Emotion Detection from Speech: An Excitation Parameter Approach” | The Iiit-H Telugu Emotional Speech Database,  The Berlin Emotional Speech Database | KL Distance Method | The main advantage of the proposed system is that there is no need for training the system with emotional speech. The feature combinations that are used in this study can also help in developing an emotion recognition system. | Being speaker-specific, it may not generalize well to unfamiliar voices, and while it can be made speaker-independent, this requires additional data.  The system’s reliance on a reference utterance in a neutral state could pose challenges if the neutral state varies significantly. | 77.33% | Leveraging excitation features and KL distance, the system requires no training with emotional speech, shows potential for improved performance in emotion detection and recognition when combining features from the excitation and vocal tract system. |
| 3 | [4] | 2019 | “Synthetic Speech for Emotion Recognition: Creating Neutral Reference Models and Improving Detection of Arousal and Valence Levels” | SEMAINE database,  Wall Street Journal based Continuous Speech recognition Corpus Phase II database | Kullback-Leibler divergence (KLD) or J-divergence | Effectively removes the lexical content of the sentence, focusing on the emotional content instead. The system has shown improvements in emotion classification evaluations, with increases in the average F-score for both arousal and valence. | The system’s performance is dependent on the quality of the synthetic speech references. If these references do not adequately capture the properties of neutral speech, it could impact the system’s ability to accurately contrast emotional cues. | --- | It demonstrates the feasibility of using advances in speech synthesis to build robust neutral reference models, leading to improved performance in emotion detection and recognition. |
| 4 | [5] | 2018 | “Bridging the Affective Gap in Speech Emotion Recognition: A Deep Convolutional Neural Network Approach with Discriminant Temporal Pyramid Matching” | Berlin dataset of German emotional speech (EMO-DB), the RML audio-visual dataset, the eNTERFACE05 audio-visual dataset, and the BAUM-1s audio-visual dataset | DCNN model called AlexNet | DCNNs for automatic feature learning, which can yield promising performance in speech emotion recognition tasks. The system also uses a Discriminant Temporal Pyramid Matching (DTPM) strategy to aggregate learned segment-level features into a global utterance-level feature representation | The system focuses on global utterance-level emotion classification, and therefore, it may not be capable of dealing with continuous dimensional emotion recognition. | 69.70% on RML  86% on eNTERFACE05 | It uses DCNNs to extract segment-level features from log Mel-spectrograms and DTPM to aggregate these features into a global utterance-level feature representation for emotion recognition. |
| 5 | [6] | 2018 | Three-Dimensional Attention-Based Convolutional Recurrent Neural Networks for Speech Emotion Recognition: An Approach Using Mel-Spectrogram with Deltas and Delta-Deltas | Interactive Emotional Dyadic Motion Capture database (IEMOCAP),  the Berlin Emotional Database (Emo-DB) | ACRNN-  CRNN to extract high level feature for SER,  Attention layer, employed to focus on emotion relevant parts and produce discriminative utterance-level representations for SER | It employs a 3-Dimensional Attention-based Convolutional Recurrent Neural Networks (3-D ACRNN) for automatic feature learning, which has shown to outperform the state-of-the-art DNN-ELM method | The system seems to have difficulty distinguishing between certain emotions, such as happy and angry, due to similar activation levels. | 82% on EmoDB  64% on IEMOCAP | The paper presents a novel method for speech emotion recognition using a 3-D attention-based convolutional recurrent neural network (ACRNN) and Mel-spectrogram with deltas and delta-deltas as input1. The method shows promising results on two benchmark datasets, IEMOCAP and Emo-DB. |
| 6 | [7] | 2023 | Enhancing Cross-Corpus Speech Emotion Recognition through Local Domain Adaptation: A Transfer Learning Approach | IEMOCAP,  Emo-DB | Six (CNN) layers,  one (RNN) layer, one attention layer and one classifier from the input layer to the output layer. All implemented in Py Torch. | Local domain adaptation studies more fine-grained data alignment, which makes domain adaptation achieve better performance | Lower accuracy of the E I task due to the smaller sample size of Emo-DB. | 82% on EmoDB  64% on IEMOCAP | he paper proposes a new framework for speech emotion recognition that uses local domain adaptation and a category-grained discrepancy to evaluate the distance between two relevant domains. This approach improves the generalization ability of the model |
| 7 | [8] | 2023 | “Enhancing Speech Emotion Recognition with Mel-STFT and Improved Multiscale Vision Transformers: A Comprehensive Approach” | Emo-DB, RAVDESS, and IEMOCAP | Gaussian Naïve Baiyes, SVM,  Random Forest,  KNN,  MLP | It outperforms existing methods on all three datasets. The mel-STFT features are effective in capturing relevant acoustic information related to speech, such as spectral characteristics and energy distribution | Potential vulnerability to misclassification of emotions due to the complex and sometimes overlapping nature of emotional expressions, | 90.75% on EmoDB  81% on RAVDESS | The paper presents a method for speech emotion recognition that combines Mel spectrogram with Short-Term Fourier Transform (Mel-STFT) and Improved Multiscale Vision Transformers (MViTv2). It provides comprehensive representation of emotional content in speech signals and achieves high accuracy on various datasets. |
| 8 | [9] | 2023 | Efficient Lightweight Model for Speech Emotion Recognition Using Random Forest and MLP Classifiers within VGGNet Framewor | Tess Dataset,  RAVDESS,  EmoDB | Random Forest,  Multi-Layer Perceptron (MLP) | proposed model is optimized for lightweight devices and has a moderate size, making it suitable for real-life applications on low-memory devices. | Even though the model achieves high accuracy on these datasets, it is unclear how well it will generalize to other datasets | 96% on EmoDB  86% on VGGNet | The paper presents a lightweight model for Speech Emotion Recognition (SER) that integrates Random Forest and Multilayer Perceptron classifiers into the VGGNet framework1. The model, which uses MFCC features, was tested on three datasets and achieved high accuracy, outperforming other deep learning methods. |

References:

[1] Shukla A., Verma S., Tiwari U.,

|  |
| --- |
| [2] W. Fan, X. Xu, B. Cai and X. Xing, "ISNet: Individual Standardization Network for Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1803-1814, 2022, doi: 10.1109/TASLP.2022.3171965.  <https://ieeexplore.ieee.org/document/9767771> |
| [3] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.  <https://ieeexplore.ieee.org/document/9046041> |
| [4] R. Lotfian and C. Busso, "Lexical Dependent Emotion Detection Using Synthetic Speech Reference," in IEEE Access, vol. 7, pp. 22071-22085, 2019, doi: 10.1109/ACCESS.2019.2898353.  <https://ieeexplore.ieee.org/document/8637936> |
| [5] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi: 10.1109/TMM.2017.2766843.  <https://ieeexplore.ieee.org/document/8085174> |
| [6] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.  <https://ieeexplore.ieee.org/document/8421023> |
| [7] Z. Huijuan, Y. Ning and W. Ruchuan, "Improved Cross-Corpus Speech Emotion Recognition Using Deep Local Domain Adaptation," in Chinese Journal of Electronics, vol. 32, no. 3, pp. 640-646, May 2023, doi: 10.23919/cje.2021.00.196.  <https://ieeexplore.ieee.org/document/10124696> |
| [8] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim and A. Alqahtani, "Mel-MViTv2: Enhanced Speech Emotion Recognition With Mel Spectrogram and Improved Multiscale Vision Transformers," in IEEE Access, vol. 11, pp. 108571-108579, 2023, doi: 10.1109/ACCESS.2023.3321122.  <https://ieeexplore.ieee.org/document/10268420>  [9] S. Akinpelu, S. Viriri and A. Adegun, "Lightweight Deep Learning Framework for Speech Emotion Recognition," in IEEE Access, vol. 11, pp. 77086-77098, 2023, doi: 10.1109/ACCESS.2023.3297269.  https://ieeexplore.ieee.org/document/10188865 |