Alex Net-Based Speech Emotion Recognition Using 3D Mel-Spectrograms

Sara Ali[1,3], Bushra Naz[2], Sanam Narejo[2,] and Zohaib Ahmed[1]
[1] Institute of Information and Communication Technologies (IICT), MUET Jamshoro
[2] Department of Computer Systems Engineering, MUET, Jamshoro.
[3] National Center of Robotics and Automation.
* **Correspondence**: Sara Ali (saratayyabali@gmail.com).

Citation | Ali. S, Naz B, Narejo S. Ahmed Z "Alex Net-Based Speech Emotion Recognition Using 3D Mel-Spectrograms", IJIST, Vol. 6 Issue. 2 pp 426-433, Apr 2024
DOI | https://doi.org/10.33411/ijasd/202462426433
Received: Mar 31, 2024 **| Revised:** April 18, 2024 **| Accepted :** April 25, 2024 **| Published :** April 27, 2024.

Speech Emotion Recognition (SER) is considered a challenging task in the domain of Human-Computer Interaction (HCI) due to the complex nature of audio signals. To overcome this challenge, we devised a novel method to fine-tune Convolutional Neural Networks (CNNs) for accurate recognition of speech emotion. This research utilized the spectrogram representation of audio signals as input to train a modified Alex Net model capable of processing signals of varying lengths. The IEMOCAP dataset was utilized to identify multiple emotional states such as happy, sad, angry, and neutral from the speech. The audio signal was preprocessed to extract a 3D spectrogram that represents time, frequencies, and color amplitudes as key features. The output of the modified Alex Net model is a 256-dimensional vector. The model achieved adequate accuracy, highlighting the effectiveness of CNNs and 3D Mel-Spectrograms in achieving precise and efficient speech emotion recognition, thus paving the way for significant advancements in this domain.

Keywords: Alex Net; Convolution Neural Network; Mel-Spectrogram; Speech Emotion Recognition.

Acknowledgement:
We greatly acknowledge the support of the National Center of Robotics and Automation and ICT endowment fund at Mehran University of Engineering and Technology, Jamshoro for funding this research.

We would like to formally acknowledge that the manuscript has not been published or submitted to other journals previously.

Author's Contribution:
Dr. Bushra Naz conceptualization and data curation; Sara Ali methodology, implementation of deep learning model, and writing; Dr. Sanam Narejo's formal analysis, reviewing, and editing; Zohaib Ahmed data preprocessing.

Conflict of Interest:
The authors declare no conflict of interest.

Introduction:

Speech Emotion Recognition (SER) involves extracting emotional features from speech signals using different classification models to recognize them [1][2]. Humans use a spectrum of different ways to express their emotions, such as body movements [3], color combinations in art, and rhythmic patterns in music [4][5]. Understanding and interpreting human emotions from audio signals is challenging due to the subjective nature of emotional content in the speech signal [6]. Traditional methods of speech emotion recognition often face limitations in capturing the nuanced features of emotions [7]. This research addresses the need for more robust and context-aware emotion recognition systems in audio signals.

Current developments in the field of audio emotion recognition focus on advanced techniques such as machine learning and signal processing, as highlighted by [8] and [9]. Multiple machine learning models like Support Vector Machine (SVM), Hidden Markov Models (HMM), Gaussian Matrix Models (GMM), and K-Nearest Neighbor [10][11] have been used for classifying the emotions within speech. Recent advancements in computing hardware and the advent of deep learning, have sparked significant interest in SER. The authors in [12] provide a comprehensive analysis of SER and propose three useful methods for identifying the emotions in speech. The study also highlighted the need to use appropriate methods for classification to improve the accuracy of the SER system. A separate investigation [13] examined Deep Learning approaches, including their features, advantages, and disadvantages. The research also divided these approaches into discriminative (Recurrent Neural Network, CNN), generative (Deep Belief Network, limited Boltzmann machine, and deep autoencoders), and hybrid categories.

However, despite the significant progress in audio signal processing, there is still a notable gap in fully understanding the complicated nature of human emotions conveyed through speech. In response to these challenges, our research adopts a structured approach to enhance audio emotion recognition. This research drew inspiration from [14] and explored the extraction of spectrogram features from audio signals. Spectrograms provide a visual representation of the frequency content of a signal over time, allowing the researchers to treat audio data as visual data and apply advanced algorithms like CNN and its derivatives from the well-established field of computer vision to analyze and interpret speech patterns.

Objectives:

The primary objectives of this research are to improve the performance of audio emotion recognition systems by leveraging the Alex Net architecture to capture complex patterns in audio spectrograms correlating with diverse emotional states. Our research distinguishes itself through the innovative integration of audio and visual modalities, treating audio signals as visual data and utilizing the Alex Net architecture as a learning algorithm. This novel approach results in the development of a system, designed to provide a more comprehensive and detailed understanding of emotions conveyed through speech. The use of spectrograms enhances the depth of data representation, enabling the model to capture subtleties that conventional audio-only methods might struggle to discern.

Novelty Statement:

The novelty of this work lies in the adaptation of Alex Net, originally designed for image-related tasks, to process audio spectrograms, thereby introducing a groundbreaking solution for emotion recognition in speech.

Material and Methods:

In this research, a system is developed to recognize four emotions from speech. The methodology of the research is shown in Figure 1.

Data Collection: In this research, we utilized the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, a publicly available benchmark dataset in the domain of public academic research concerned with speech and emotion. It is published by the University of

Southern California. The dataset has seven emotional categories; however, the scope of this research is only limited to four categories, i.e., neutral, happy, angry, and sad. The rationale for selecting the IEMOCAP dataset lies in its rich emotional annotations, comprising recordings of naturalistic dyadic interactions that are unscripted and spontaneous, thereby mirroring real-life conversational scenarios. Additionally, the dataset contains interactions among multiple speakers of different genders across various sessions, thereby introducing variability in speech patterns, accents, and speaking styles. This diversity is pivotal for training models capable of generalizing effectively to unseen speakers and a spectrum of diverse speech characteristics.

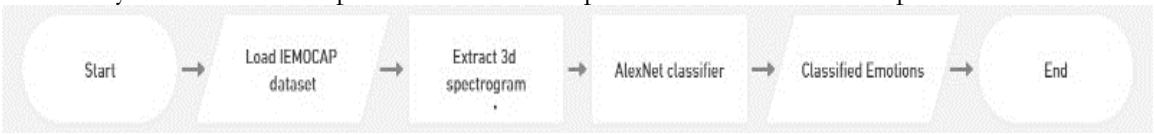

Figure 1: Flow of Methodology

In the original corpus, the instances for happy class were relatively low which created a class imbalance problem. To deal with this problem, the data of the excited class was also merged with the happy class. This not only balanced the dataset but also made our work consistent with previous studies using the same dataset [15].

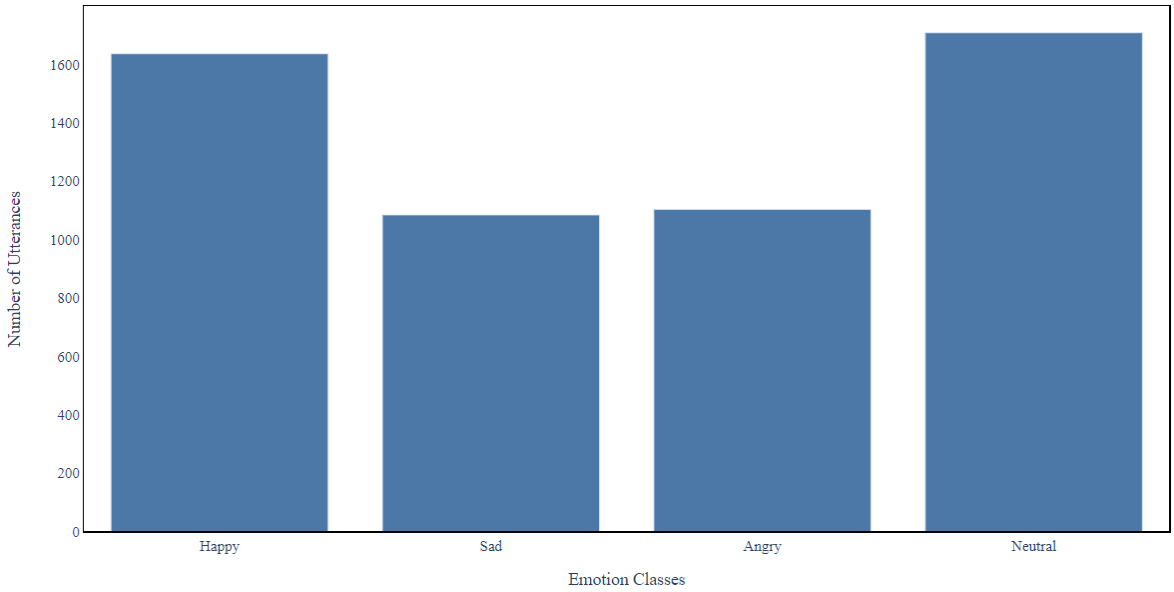

Figure 2: Distribution of Utterances across Classes

The updated dataset comprises 5,531 utterances, distributed as follows: 1636 instances of happy class, 1084 of sad class, 1103 of angry, and 1708 of neutral class, as shown in Figure 2. Out of these 4,978 samples were used for training the deep learning model whereas 553 samples were used for testing.

Feature Extraction:

The feature extraction process began with the audio data in the **.wav** format. It was gathered from the processed files along with their labels. The next step was to extract features from the audio data. Our feature extraction process was specifically focused on utilizing a 3-dimensional log Mel-spectrogram. This spectrogram representation provided a comprehensive view of the audio signals, capturing both the frequency and intensity aspects. The Mel-spectrogram was derived from the Short-Time Fourier Transform (STFT) of the audio signal, with the third dimension representing the intensity of the signal at each time-frequency bin [16]. Hence, the .wav data was converted into a 3D spectrogram using the equation below. The X-axis represents time, whereas the Y-axis is used to show the frequencies of the audio signals, and the colors are used to show the amplitude of the audio signal frequency at a particular instance [17].

$$S(n) = \sum_{x-f(m-1)}^{f(n+1)} \log(H_n(x) * |X(x)|^2)$$

The above formula was used to calculate the 3D log spectrogram of a signal using a fast Fourier transform. Where, the $|X(x)|^2$ shows the range of energy in the x^{th} position, n shows the number of filter banks, and x refers to the point or position of the Fast Fourier Transform. Spectrograms were extracted from 25 ms frames at a 100 Hz frame rate (i.e., every 10 ms). Figure 2 shows the waveforms and 3-D Log-Mel spectrograms of the Neutral, Angry, Sad, and Happy emotional states respectively in the IEMOCAP database.

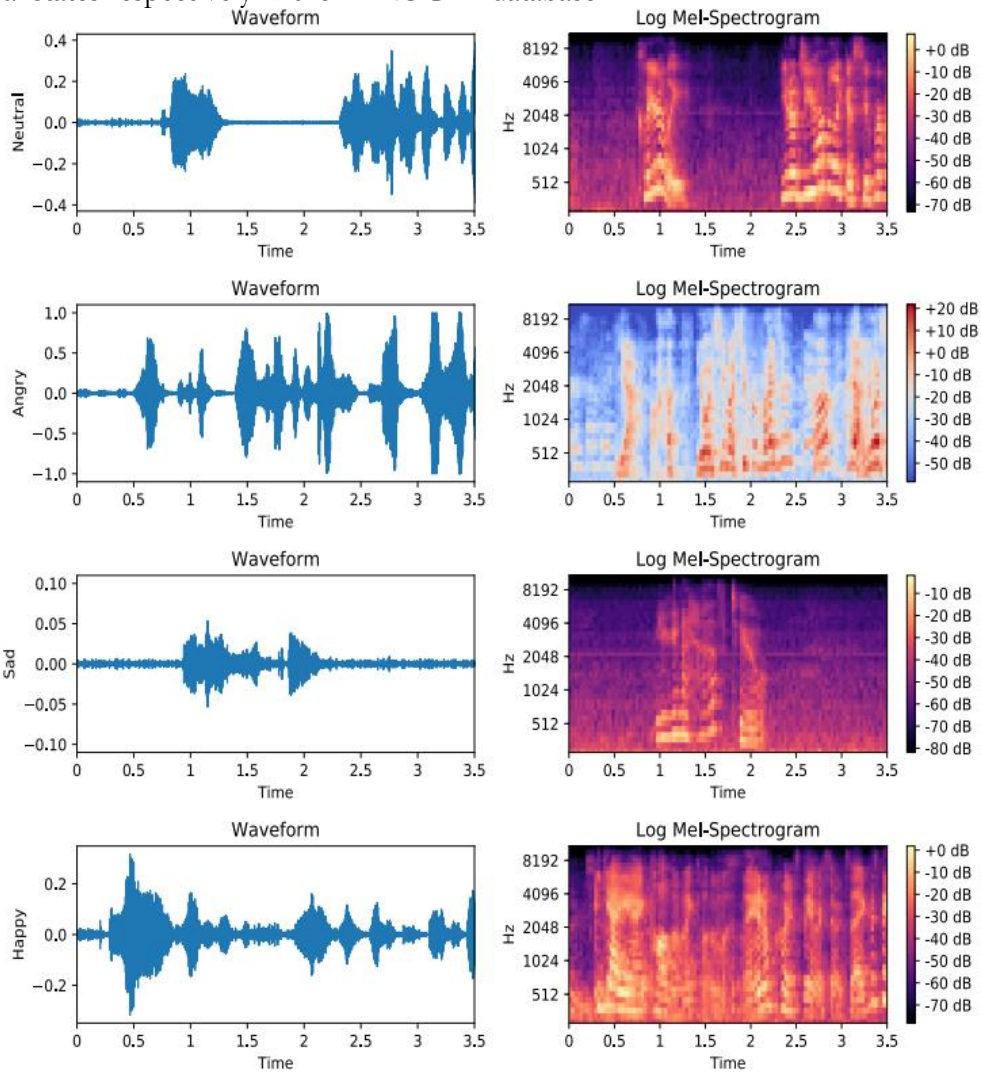

Figure 3: Waveforms and 3-D Log-Mel spectrograms of the four emotional states respectively in the IEMOCAP database [18].

The choice of using Mel-spectrogram for feature extraction is pivotal in characterizing speech signals based on their frequency content and intensity variations. The Mel-spectrogram has proven to be effective in capturing essential acoustic features related to human speech and emotions [19][20]. By concentrating on the 3-dimensional Mel-spectrogram as our primary feature, we aimed to extract rich information that was relevant for recognizing emotional nuances in the audio signals. This feature will serve as the input to our model for further analysis and emotion recognition tasks.

Alex Net Model:

The base Alex Net architecture is as follows: It has a total of eight layers including five convolutional layers (Conv1 to Conv5) and three fully connected layers (FC6, FC7, and FC8), as also shown in Figure 4. In the context of speech recognition, the input to the Alex Net model is typically 3D Mel-spectrograms, which are visual representations of audio signals over time. The input layer handles these inputs and passes them to the first convolutional layer. The five convolutional layers of Alex Net were designed to capture hierarchical features from the input data. They used filters (also known as kernels) to extract features such as edges, textures, and other patterns from the input. Each convolutional layer was followed by a Rectified Linear Unit (ReLU) activation function, which introduces non-linearity and helps the model learn complex patterns [21]. The equation for ReLU is given as:

$$f(x) = \max(0, x)$$

Local Response Normalization (LRN) was applied after the first convolutional layer to enhance feature discrimination. Max-pooling layers were used to downsample spatial dimensions, which leads to translation invariance. The architecture includes three fully connected layers (FC6, FC7, and FC8), culminating in a SoftMax activation function to convert the network's raw output into probability scores for each class. This was used for classification purposes, providing the probabilities for each class.

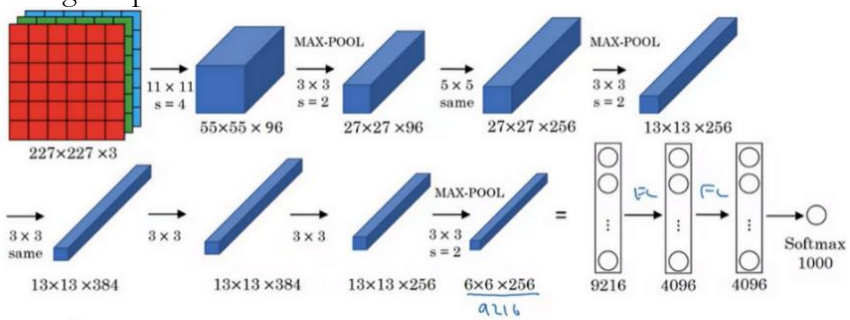

Figure 4: Alex Net architecture

Alex Net for SER:

Alex Net accepts 3D Mel-spectrograms as input, which are visual representations of audio signals over time, with frequency and amplitude information. The convolutional layers extracted important features from the spectrograms, which aided in identifying different classes of speech or emotions. The architecture was modified to handle variable-length audio signals, allowing the network to manage audio inputs of different durations. Techniques such as resizing or cropping were employed to process inputs of different dimensions. Along with that, in order to reduce over-fitting, the fully connected layer before the final SoftMax layer was removed from the architecture, making the network shallower. A shallower network is more robust to unknown examples and helps improve generalization.

Performance Indicators:

Performance indicators prove effective in order to evaluate a learning model for a classification problem, such as speech emotion recognition. The indicators/metrics utilized in this paper included F1 score, recall, precision, accuracy, and support. Support values denote the count of real instances of a particular class within the dataset. The equations for other indicators are as follows:

- **Precision:** Number of True Positives (TP) divided by the total number of True Positives (TP) and False Positives (FP).

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Number of True Positives (TP) divided by the total number of True Positives (TP) and False Negatives (FN).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1 Score:** The F1 score is the harmonic mean of precision and recall.

$$\text{F1 Score} = 2\ \text{X} \frac{\text{Precision X Recall}}{\text{Precision} + \text{Recall}}$$

However, these formulas are limited to binary classification problems where the data only has two classes. These indicators alone fall short of painting a complete picture of scenarios where multiple target classes are present, such as in this case. In such cases, averaging methods such as, weighted average and macro average help understand the performance of the model.

- **Macro Average:** The macro-averaged F1 score was computed by taking the arithmetic mean of all the per-class F1 scores. This method treats all classes equally regardless of their support values.
- **Weighted Average:** Weighted average calculates the average of a metric across each class but weighs each class by its support (the number of instances in each class). This accounts for class imbalance in the dataset.

$$\text{Weighted Average} = \sum_{i=1}^{n} (\text{Metric}_i\ \text{X}\ \frac{\text{Support}_i}{\text{Total Support}_i})$$

Result and Discussion:

The results were validated using the IEMOCAP dataset, comprising 553 instances to evaluate the model. Table 1 shows a detailed classification report of the final model with precision, recall, support, and F1 scores for each class.

Table 1: Classification report of the model

Class	Recall	F-1 Score	Precision	Support
Angry	0.82	0.80	0.73	106
Happy	0.72	0.73	0.72	162
Sad	0.72	0.69	0.66	107
Neutral	0.60	0.63	0.66	178
Accuracy	-	0.71	-	553
Average (macro)	0.71	0.72	0.71	553
Average (weighted)	0.70	0.70	0.70	553

Table 1 presents a detailed evaluation of the model's performance across different emotion classes. The precision for the Angry class was 0.77, indicating that out of all instances predicted as the True class by the model, 77% of them are actually Angry. The number of actual instances in the dataset that belong to the Angry class was 106 and out of all the actual instances 81% of them are identified correctly. The F1 Score for the Angry class was 0.79, illustrating the harmonic mean of precision and recall for the Angry class. These metrics show that the model performed reasonably well in identifying instances of the "Angry" class, with a good balance between precision and recall.

Moving to the Happy class, the model showed evidence of precision of 0.71, indicating that there is room for improvement in accurately identifying happiness. However, the balanced recall of 0.71 and F-1 score of 0.72 indicated the model's effectiveness in capturing instances of happiness. The support value of 162 ensured a sufficient dataset for comprehensive evaluation. In the Sad class, the model achieved a precision of 0.65, indicating a moderate level of accuracy. The relatively high recall of 0.71 suggested the model's effectiveness in identifying actual instances of sadness. The balanced F-1 score of 0.68 further signifies a reasonable trade-off between precision and recall. With a support value of 107, there was adequate data for meaningful evaluation.

For the Neutral class, the model demonstrated a moderate precision of 0.65 and a balanced recall of 0.59, indicating its capability to correctly predict neutrality. The F-1 score of 0.62 reflected a fair trade-off between precision and recall. The support value of 178 contributes to reliable evaluation. The model showed an overall accuracy of 0.70, showcasing its correctness in classifying instances across all emotion classes. Macro averages (0.70, 0.70, 0.71) and weighted averages (0.69, 0.69, 0.69) provided a comprehensive assessment, treating each class equally and considering class imbalances, respectively. However, it's important to note that challenges such as class imbalance and data variability could impact the generalization of the model to unseen data. A key limitation of this study is that it only focuses on the utilization of audio for SER. This opens an opportunity for future research direction where multiple modalities like text and videos could also be used along with the audio spectrograms.

Conclusion:

In conclusion, SER presents a challenging task primarily due to the intricate nature of the audio signals. To deal with this problem, this research demonstrated the effectiveness of a modified Alex Net model with 3D Mel-Spectrograms for speech emotion recognition. This approach allowed the audio data to be represented in visual form with the help of 3D spectrograms. The model performed reasonably well in predicting anger and demonstrated the ability to recognize anger, happiness, sadness, and neutrality. Areas for potential improvement lie particularly in precision for happiness, declaring further refinement through additional data or fine-tuning. Investigating misclassifications and enhancing model interpretability may offer valuable insights for future enhancements.

References:

[1] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," IEEE Access, vol. 9, pp. 47795–47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[2] Mustaqeem and S. Kwon, "CLSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network," Math. 2020, Vol. 8, Page 2133, vol. 8, no. 12, p. 2133, Nov. 2020, doi: 10.3390/MATH8122133.

[3] M. M. N. Bieńkiewicz et al., "Bridging the gap between emotion and joint action," Neurosci. Biobehav. Rev., vol. 131, pp. 806–833, Dec. 2021, doi: 10.1016/J.NEUBIOREV.2021.08.014.

[4] V. Chaturvedi, A. B. Kaur, V. Varshney, A. Garg, G. S. Chhabra, and M. Kumar, "Music mood and human emotion recognition based on physiological signals: a systematic review," Multimed. Syst., vol. 28, no. 1, pp. 21–44, Feb. 2022, doi: 10.1007/S00530-021-00786-6/METRICS.

[5] "Is happier music groovier?: The influence of emotional characteristics of musical chord progressions on groove." Accessed: Apr. 14, 2024. [Online]. Available: https://osf.io/preprints/osf/h3wrm

[6] A. I. Iliev and A. I. Iliev, "Perspective Chapter: Emotion Detection Using Speech Analysis and Deep Learning," Emot. Recognit. - Recent Adv. New Perspect. Appl., May 2023, doi: 10.5772/INTECHOPEN.110730.

[7] S. Islam, M. M. Haque, and A. J. M. Sadat, "Capturing Spectral and Long-term Contextual Information for Speech Emotion Recognition Using Deep Learning Techniques," Aug. 2023, Accessed: Apr. 14, 2024. [Online]. Available: https://arxiv.org/abs/2308.04517v1

[8] M. Dong, L. Peng, Q. Nie, and W. Li, "Speech Signal Processing of Industrial Speech Recognition," J. Phys. Conf. Ser., vol. 2508, no. 1, p. 012039, May 2023, doi: 10.1088/1742-6596/2508/1/012039.

[9] N. Bashir et al., "A Machine Learning Framework for Major Depressive Disorder (MDD) Detection Using Non-invasive EEG Signals," Wirel. Pers. Commun., pp. 1–23,

May 2023, doi: 10.1007/S11277-023-10445-W/METRICS.
[10] R. Jahangir, Y. W. Teh, G. Mujtaba, R. Alroobaea, Z. H. Shaikh, and I. Ali, "Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion," Mach. Vis. Appl., vol. 33, no. 3, pp. 1–16, May 2022, doi: 10.1007/S00138-022-01294-X/METRICS.
[11] R. H. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic Speech Emotion Recognition from Saudi Dialect Corpus," IEEE Access, vol. 9, pp. 127081–127085, 2021, doi: 10.1109/ACCESS.2021.3110992.
[12] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Commun., vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/J.SPECOM.2019.12.001.
[13] R. Jahangir, Y. W. Teh, F. Hanif, and G. Mujtaba, "Correction to: Deep learning approaches for speech emotion recognition: state of the art and research challenges (Multimedia Tools and Applications, (2021), 80, 16, (23745-23812), 10.1007/s11042-020-09874-7)," Multimed. Tools Appl., vol. 80, no. 16, p. 23813, Jul. 2021, doi: 10.1007/S11042-021-10967-0/METRICS.
[14] "Decoding the Symphony of Sound: Audio Signal Processing for Musical Engineering | by Naman Agrawal | Towards Data Science." Accessed: Apr. 14, 2024. [Online]. Available: https://towardsdatascience.com/decoding-the-symphony-of-sound-audio-signal-processing-for-musical-engineering-c66f09a4d0f5
[15] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang. Resour. Eval., vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/S10579-008-9076-6/METRICS.
[16] Y. Wang, H. Zhang, and W. Huang, "Fast ship radiated noise recognition using three-dimensional mel-spectrograms with an additive attention based transformer," Front. Mar. Sci., vol. 10, no. November, pp. 1–14, 2023, doi: 10.3389/fmars.2023.1280708.
[17] A. Rodriguez, Y. L. Chen, and C. Argueta, "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," IEEE Access, vol. 10, pp. 22400–22419, 2022, doi: 10.1109/ACCESS.2022.3151098.
[18] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network," IEEE Access, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.
[19] S. Madanian et al., "Speech emotion recognition using machine learning — A systematic review," Intell. Syst. with Appl., vol. 20, p. 200266, Nov. 2023, doi: 10.1016/J.ISWA.2023.200266.
[20] I. Pulatov, R. Oteniyazov, F. Makhmudov, and Y. I. Cho, "Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders," Sensors 2023, Vol. 23, Page 6640, vol. 23, no. 14, p. 6640, Jul. 2023, doi: 10.3390/S23146640.
[21] A. K. Dubey and V. Jain, "Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions," Lect. Notes Electr. Eng., vol. 553, pp. 873–880, 2019, doi: 10.1007/978-981-13-6772-4_76/COVER.