

Real-Time Speech Emotion Recognition Using Lightweight Convolutional Neural Network

Natchathiram K¹, Prof. Dr. G. Kulanthaivel², Prof. Dr. P. Sivasankar³

^{1, 2, 3}Dept of Electronics and Communication Engineering,

^{1, 2, 3}National Institute of Technical Teachers Training and Research (NITTTR), Chennai, India

Abstract- *Speech Emotion Recognition (SER) allows machines to understand human emotional states from their speech, enabling applications in human-computer interaction, call center analytics, and mental health monitoring. Traditional deep CNNs achieve high accuracy; however, they are costly when it comes to computation, which means they cannot be used in real time on edge devices with limited computational power. This paper proposes a lightweight CNN-based architecture for real-time SER using Log-Mel spectrograms as input representations. The proposed model incorporates Conv2D layers, batch normalization, max pooling, and global average pooling to learn discriminative time-frequency features while reducing model complexity and inference overhead. Unlike the existing Artificial Neural Network (ANN) baseline, which was trained using only the RAVDESS dataset, the proposed CNN is trained using a combined RAVDESS and TESS dataset covering eight emotion classes: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Experimental evaluation shows that the proposed Lightweight CNN attained 75.70% test accuracy with a weighted F1-score of 0.78, compared to the ANN baseline model's 54.86% accuracy and 0.55 F1-score. Additionally, compared to the ANN baseline, the proposed Lightweight CNN architecture demonstrated faster real-time performance using microphone input while providing confidence scores for each emotion detected. Overall, the proposed lightweight CNN model provides faster inference with good accuracy for SER applications.*

Keywords- Speech Emotion Recognition, Lightweight CNN, Real-Time Systems, Log-Mel Spectrogram, RAVDESS, TESS

I. INTRODUCTION

Speech signal, being a physiological process, captures the change in the emotional state of the speaker. The changes in speech can indicate a person's identity, mental, and physical health status [1]. Detecting the emotional state of the speaker from speech signals refers to a research field termed speech emotion recognition (SER) [2]. SER plays a vital role in human-computer interaction (HCI). The growing advancements in SER are largely driven by its significance in HCI. Humans naturally perceive emotions through speech, while robust SER aims to equip machines with a similar

ability to accurately recognize emotions from speech utterances [3]. SER finds applications in customer service monitoring, virtual assistants, e-learning, psychological health assessment, and safety-critical systems like driver fatigue detection [4].

SER system performs two tasks, namely feature extraction and classification/recognition. The feature extraction task is important since the performance of SER mainly depends on it. The extracted features must have the ability to represent different emotions from speech signals, and at the same time they must be invariant to speaker characteristics. In this paper, a lightweight CNN architecture is proposed using Log-Mel spectrogram representations. The objective is to develop a computationally efficient model while maintaining reliable emotion classification performance. Data augmentation is incorporated to increase the diversity of training samples and improve robustness against variations in speech signals.

The major contributions of this work are as follows:

- A complete SER framework using RAVDESS and TESS datasets is developed.
- Multiple audio preprocessing and augmentation techniques are employed to improve the robustness of the training data.
- A fixed 64×128 Log-Mel spectrograms representation are used as CNN input features.
- A lightweight CNN architecture is developed using convolution, batch normalization, max pooling, global average pooling, and dense layers.
- The system supports eight emotion classes and can be extended toward real-time speech emotion prediction.

The remainder of this paper is organized as follows: Section II presents related SER work; Section III explains the proposed methodology and the CNN architecture; Section IV describes the experimental setup; Section V discusses the obtained results. Finally, Section VI concludes the work and future improvements.

II. RELATED WORK

Various feature extraction techniques have been investigated, including spectral features, cepstral features, prosodic features, wavelet features [8]-[14]. Based on these representations, different classification models have been explored, such as support vector machines, artificial neural networks, Gaussian mixture models, hidden Markov models, k-nearest neighbors, XGBoost (eXtreme Gradient Boosting) [2], [14]-[18].

Aljuhani [6] presented a speech emotion recognition based on spoken data in Saudi Arabian Arabic dialectal using speech collected from freely available YouTube videos and consisted of four types of emotions (anger, happiness, sadness, and neutral). MFCCs and Mel spectrograms, were extracted. These features were then used as inputs for the KNN classifier. This experiments gives an overall accuracy of 57.14% using KNN.

Bhandari [11] evaluated an approach based on LSTM network models as a classifier using the MFCC features extracted from the RAVDEES dataset. Only 4 emotion categories were employed (happy, neutral, sad, and angry). This approach achieved low accuracy. They achieved a recognition rate of 66%.

Singh [15] introduced a Constant-Q Transform with Modulation Spectral Feature (CQT-MSF), for Speech Emotion Recognition (SER). Implementation is done through 2-dimensional convolutional networks that give time-shift deformation as well as invariant features related to SER.

Hama Saeed [17] presented a DNN based SER approach. This approach is divided into three stages: feature extraction, normalization, and emotion recognition. This study considers MFCC, Mel- Spectrogram Frequency, Chroma, and Poly Features were extracted from the audio signals. Data augmentation for the minority class using SMOTE and the Min-Max scaler for the normalization process were used. However, comparatively fewer studies have systematically applied and evaluated such lightweight design principles specifically for real-time speech emotion recognition, which motivates the design proposed in this paper.

III. DETAILS METHODOLOGY

The proposed lightweight Convolutional Neural Network (CNN)-based Speech Emotion Recognition (SER) framework is designed to identify human emotions from speech signals with reduced computational complexity. The system uses the RAVDESS and TESS datasets for model training and supports real-time emotion prediction using

microphone input. The complete processing pipeline consists of speech input, preprocessing, data augmentation, Log-Mel spectrogram generation, CNN-based feature extraction, classification, and emotion prediction as illustrated in Figure 1.

A. Speech Input

The proposed system accepts speech signals from two sources: publicly available emotional speech datasets and live microphone input. The proposed system, in contrast, is trained using a combination of two benchmark datasets — RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotional Speech Set) — together with live microphone input for real-time inference. Combining RAVDESS and TESS increases both the volume and speaker/style diversity of the training data (RAVDESS contributes professional actor recordings with song and speech, while TESS contributes controlled recordings from two female speakers across the same emotion set), which supports better generalization in the proposed Lightweight CNN. The system considers eight emotion classes: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised.

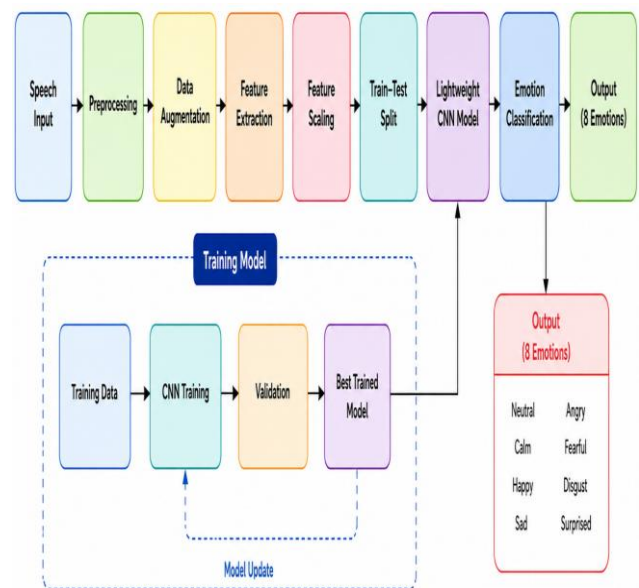


Figure 1. Proposed system architecture for real-time speech emotion recognition using a Lightweight CNN

B. Audio Preprocessing

The raw speech signals may contain silence, amplitude variations, and differences in sampling rates. Therefore, preprocessing is performed before feature extraction. The preprocessing stage includes: Audio Loading (The speech files are loaded using a fixed sampling rate), Silence Removal (Silent portions at the beginning and end of

the recording are removed), Normalization (The amplitude of the speech signal is normalized to reduce variations between recordings), and Resampling (Audio signals are converted to a common sampling rate to maintain consistency). These operations produce a standardized speech signal suitable for subsequent processing.

C. Data Augmentation

To improve model robustness and reduce overfitting, data augmentation is applied to the training samples. The proposed system uses controlled variations of the original speech signals without changing their corresponding emotion labels.

The augmentation techniques include:

1) Noise Addition

Small-amplitude random noise is added to the original speech signal. This improves robustness against background noise and recording variations.

$$x_{aug}(n) = x(n) + 0.005 \eta(n)$$

2) Time Shifting

The speech signal is shifted along the time axis. This introduces temporal variation while preserving the emotional characteristics of the speech.

3) Pitch Shifting

The pitch of the speech signal is slightly modified to introduce speaker-level variability. This helps the model learn emotion-related patterns rather than depending on a particular speaker's pitch characteristics.

Thus, augmentation increases the diversity of the training samples and improves the generalization capability of the proposed model.

D. Feature Extraction and Scaling

The proposed system converts each processed speech signal into a two-dimensional Log-Mel spectrogram. Each spectrogram is fixed to a feature size of 64×128 (Mel bands $\times$ time frames), giving a compact yet emotion-rich time-frequency representation suitable as input to the CNN.

Every audio sample is converted into a fixed-size two-dimensional representation before being supplied to the CNN.

Feature scaling is applied to reduce variations in feature magnitude. This process helps the neural network receive features with a consistent numerical distribution and can improve training stability.

E. Train-Test Data Splitting

After preprocessing, augmentation, and feature extraction, the complete dataset is divided into training and testing subsets.

An 80:20 split is employed: Training data: 80% and Testing data: 20%

The training data is used for learning the CNN parameters, whereas the testing data is kept separate for evaluating the generalization capability of the trained model.

To avoid data leakage, augmented versions of an audio sample should ideally remain within the same partition as the corresponding original sample.

F. Lightweight CNN Architecture

The proposed model uses a lightweight CNN architecture as shown in Figure 2. to reduce computational complexity while extracting important time-frequency patterns from the Log-Mel spectrogram.

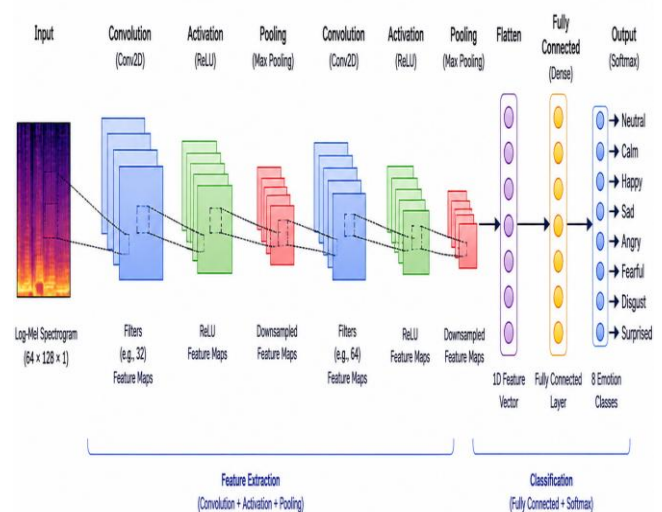


Figure 2. Lightweight CNN Architecture

The main components of the architecture are:

1. Convolutional Layer

The convolutional layers learn local spectral and temporal patterns from the Log-Mel spectrogram. Different convolution kernels detect different patterns associated with speech characteristics.

2. ReLU Activation

The Rectified Linear Unit (ReLU) activation function introduces non-linearity:

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

This enables the network to learn complex relationships between acoustic patterns and emotional categories.

3. Batch Normalization

Batch normalization is used to normalize intermediate feature representations during training. It helps stabilize the learning process and can improve convergence.

4. Max Pooling

Max pooling reduces the spatial dimensions of feature maps while retaining dominant activation responses. This reduces computational requirements and provides a degree of translation invariance.

5. Global Average Pooling

Global Average Pooling (GAP) converts the final feature maps into a compact feature vector by calculating the average activation of each feature map.

Compared with large fully connected layers, GAP significantly reduces the number of trainable parameters.

6. Dense Layer

The extracted high-level features are passed to a dense layer for final classification processing.

7. Softmax Classifier

The final layer contains eight output neurons corresponding to the eight emotion categories. Softmax converts the output logits into class

probabilities:

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^8 e^{z_j}}$$

The emotion with the highest probability is selected as the predicted emotion.

The CNN is trained using the augmented training dataset. During training, the network learns the relationship between the Log-Mel spectrogram patterns and the corresponding emotion labels.

Table 1. Proposed Lightweight CNN architecture (representative configuration)

Layer	Type	Output Shape	Parameters
1	Input (Log-Mel Spectrogram)	$64 \times 128 \times 1$	0
2	Conv2D (32 filters, 3×3) + ReLU	$64 \times 128 \times 32$	~320
3	Batch Normalization	$64 \times 128 \times 32$	~128
4	Max Pooling (2×2)	$32 \times 64 \times 32$	0
5	Conv2D (64 filters, 3×3) + ReLU	$32 \times 64 \times 64$	~18.5K
6	Batch Normalization	$32 \times 64 \times 64$	~256
7	Max Pooling (2×2)	$16 \times 32 \times 64$	0
8	Conv2D (128 filters, 3×3) + ReLU	$16 \times 32 \times 128$	~73.9K
9	Batch Normalization	$16 \times 32 \times 128$	~512
10	Global Average Pooling	128	0
11	Dense + Softmax (Output)	8 (emotion classes)	~1K

The proposed model performs eight-class emotion classification. For an input speech signal, the CNN generates a probability distribution across all eight classes. The class with the maximum probability is selected as the final predicted emotion. A representative layer configuration is summarized in Table 1.

IV. EXPERIMENTAL SETUP

The proposed system is implemented using Python and commonly used audio-processing and deep-learning libraries. Experiments are conducted on a system with [specify GPU/CPU, RAM]. Hyperparameters, including batch size, learning rate, and number of training epochs, are tuned via validation performance; representative values are batch size = 32, initial learning rate = $1e-3$, and a maximum of 30 training epochs with early stopping. The lightweight architecture is selected with the objective of reducing computational requirements compared with larger CNN architectures.

V. RESULTS AND DISCUSSIONS

The proposed lightweight CNN is evaluated using the unseen emotional speech dataset. The existing system, based on a standard ANN architecture, was trained and evaluated using only the RAVDESS dataset, with a dataset shape of (1440, 64). The model achieved a training accuracy of 91.67% but a considerably lower test accuracy of 54.86%, indicating a large generalization gap and overfitting to the training data.

The class-wise classification report using ANN is presented in Table 2.

Table 2. Classification report — Existing ANN model

Emotion	Precision	Recall	F1-Score	Support
Angry	0.81	0.58	0.68	36
Calm	0.55	0.59	0.57	27
Disgust	0.67	0.40	0.50	35
Fearful	0.57	0.68	0.62	53
Happy	0.48	0.54	0.51	37
Neutral	0.24	0.35	0.29	17
Sad	0.52	0.36	0.42	45
Surprised	0.57	0.76	0.65	38
Accuracy			0.55	288
Macro avg	0.55	0.53	0.53	288
Weighted avg	0.57	0.55	0.55	288

The proposed Lightweight CNN was trained on the combined RAVDESS and TESS datasets using Log-Mel spectrogram features with data augmentation, and validated in real time using live microphone input. It achieved a test accuracy of 75.70% with a test loss of 0.6175 — a substantial improvement over the ANN baseline. Table 3 reports the per-class precision, recall, and F1-score.

The proposed model attains strong per-class performance on neutral (F1 = 0.91), angry (F1 = 0.86), and disgust (F1 = 0.83), and shows a marked improvement over the ANN baseline across nearly every class. Table 4 summarizes the overall performance comparison between the existing ANN model and the proposed Lightweight CNN.

Table 3. Classification report — Proposed Lightweight CNN

Emotion	Precision	Recall	F1-Score	Support
Angry	0.94	0.80	0.86	592
Calm	0.18	0.52	0.26	192
Disgust	0.80	0.87	0.83	592
Fearful	0.85	0.81	0.83	592
Happy	0.95	0.70	0.81	592
Neutral	0.99	0.84	0.91	496
Sad	0.80	0.74	0.77	592
Surprised	0.46	0.38	0.41	192
Accuracy			0.76	3840
Macro avg	0.75	0.71	0.71	3840
Weighted avg	0.83	0.76	0.78	3840

Table 4. Overall performance comparison — Existing ANN vs. Proposed Lightweight CNN

Metric	Existing ANN Model	Proposed Lightweight CNN
--------	--------------------	--------------------------

Training Dataset	RAVDESS only	RAVDESS + TESS
Train Accuracy	91.67%	70.20%(final epoch)
Test / Validation Accuracy	54.86%	75.70%
Test Loss	—	0.6175
Weighted F1-Score	0.55	0.78
Macro F1-Score	0.53	0.71
Generalization Gap (Train – Test)	36.81% (overfitting)	≈ -5.5% (validation ≥ train)

The proposed CNN achieved a test accuracy of 75.70% with a test loss of 0.6175. The result indicates that the model can effectively distinguish the majority of the eight emotional categories from unseen speech samples. Figures 3 and 4 show the training/validation accuracy and loss curves respectively.

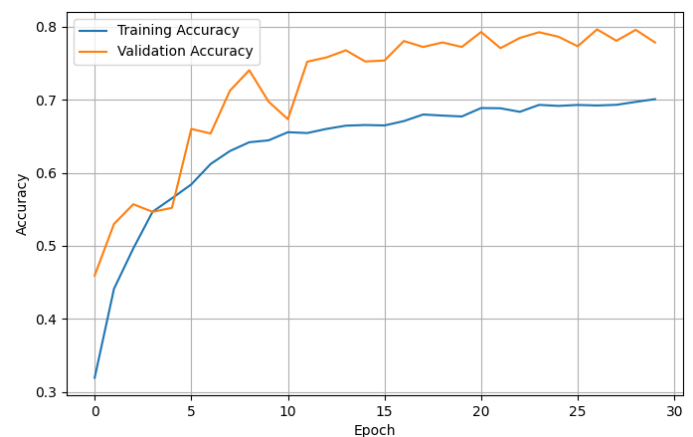


Figure 3. Training and validation accuracy

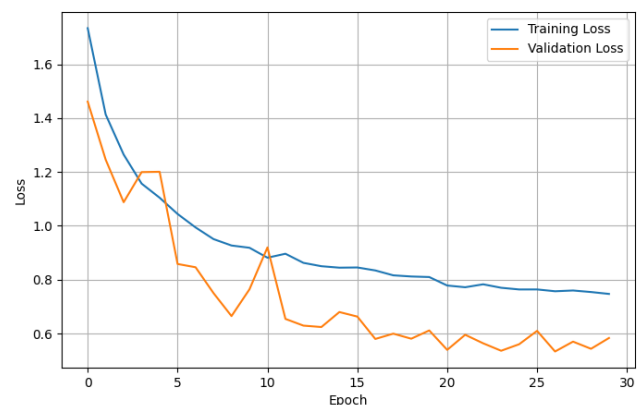


Figure 4. Training and validation loss

The trained CNN model was further validated through two real-time inference demonstrations: (i) prediction on a held-out dataset audio sample, and (ii) prediction on live microphone input captured directly from a user. For the

dataset sample, the model predicted the emotion happy with a confidence of 100.00%, with all other class probabilities at 0.00%, demonstrating a highly confident and correct prediction on a clean, in-distribution sample. For the live microphone recording, the system correctly captured audio via a record-on-Enter interface and predicted the emotion sad with a confidence of 36.04%, followed by calm (33.86%) and neutral (19.92%), with the remaining classes below 5%. Table 5 summarizes both prediction instances.

Table 5. Real-time inference results on dataset sample and live microphone input

Input Source	Predicted Emotion	Confidence	Next-Highest Class
Dataset sample	Happy	100.00%	— (all others 0.00%)
Live microphone	Sad	36.04%	Calm (33.86%), Neutral (19.92%)

VI. CONCLUSION AND FUTURE WORK

This paper presented a Lightweight CNN architecture for real-time Speech Emotion Recognition, designed using Conv2D blocks with batch normalization and global average pooling to minimize computational overhead while maintaining strong classification accuracy. Compared against an existing ANN baseline, the proposed Lightweight CNN model improved test accuracy from 54.86% to 75.70% and weighted F1-score from 0.55 to 0.78, while also showing substantially reduced overfitting, as evidenced by validation accuracy tracking closely with training accuracy across 30 epochs. Real-time validation on both a dataset sample and live microphone input confirmed the feasibility of deploying the proposed system for real-time, on-device emotion recognition.

Future work includes extending the model with attention mechanisms to better capture temporally salient emotional cues, applying quantization and pruning for further compression, improving robustness on the calm–sad–neutral and surprised–fearful confusion pairs through targeted augmentation, evaluating cross-corpus and cross-lingual generalization, and integrating multimodal cues (e.g., facial expression) to improve robustness in real-world, noisy environments.

REFERENCES

- [1] T. Özseven, “Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition,” *Appl. Acoust.*, vol. 142, pp. 70–77, Dec. 2018.
- [2] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.
- [3] S. Zhao, F. Rudzicz, L. G. Carvalho, C. Marquez-Chin, and S. Livingstone, “Automatic detection of expressed emotion in Parkinson’s disease,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4813–4817.
- [4] T. Kostoulas, I. Mporas, O. Kocsis, T. Ganchev, N. Katsaounos, J. J. Santamaria, S. Jimenez-Murcia, F. Fernandez-Aranda, and N. Fakotakis, “Affective speech interface in serious games for supporting therapy of mental disorders,” *Expert Syst. Appl.*, vol. 39, no. 12, pp. 11072–11079, Sep. 2012.
- [5] Abbaschian BJ, Sierra-Sosa D, Elmaghraby A (2021) Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21(4):1249
- [6] Abdelhamid AA, El-Kenawy E-SM, Alotaibi B, Amer GM, Abdelkader MY, Ibrahim A, Eid MM (2022) Robust speech emotion recognition using cnn+ lstm based on stochastic fractal search optimization algorithm. *IEEE Access* 10:49265–49284
- [7] Abdullah SMS, Abdulazeez AM (2021) Facial expression recognition based on deep learning convolution neural network: a review. *J Soft Comput Data Mining* 2(1):53–65
- [8] Aggarwal A, Srivastava N, Singh D (2022) Alnuaim: two-way feature extraction for speech emotion recognition using deep learning. *Sensors* 22(6):2378
- [9] Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun* 116:56–76
- [10] Aljuhani RH, Alshutayri A, Alahdal S (2021) Arabic speech emotion recognition from Saudi dialect corpus. *IEEE Access* 9:127081–127085
- [11] Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS (2023) Speech emotion recognition through hybrid features and convolutional neural network. *Appl Sci* 13(8):4750
- [12] Alnuaim, Hatamleh (2022) Human–computer interaction for recognizing speech emotions using multilayer perceptron classifier, vol. 2022. Hindawi
- [13] Aouani H, Ben Ayed Y (2020) Speech emotion recognition with deep learning. *Proc Comput Sci* 176:251–260
- [14] Arguel A, Lockyer L, Kennedy G, Lodge JM, Pachman M (2019) Seeking optimal confusion: a review on epistemic emotion management in interactive digital learning environments. *Interact Learn Environ* 27(2):200–210
- [15] Bhandari SU, Kumbhar HS, Harpale VK, Dhamale TD (2022) On the evaluation and implementation of lstm model for speech emotion recognition using mfcc. In: *Proceedings of international conference on computational*

- intelligence and data Engineering: ICCIDE 2021, pp. 421–434. Springer
- [16] Bhangale K, Kothandaraman M (2023) Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics* 12(4):839
- [17] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B et al (2005) A database of German emotional speech. *Interspeech* 5:1517–1520
- [18] Fabian Benitez-Quiroz C, Srinivasan R, Martinez AM (2016) Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570
- [19] Garg U, Agarwal S, Gupta S, Dutt R, Singh D (2020) Prediction of emotions from the audio speech signals using mfcc, mel and chroma. In: *2020 12th international conference on computational intelligence and communication networks (CICN)*, pp. 87–91. IEEE
- [20] Gupta D, Bansal P, Choudhary K (2018) The state of the art of feature extraction techniques in speech recognition. *Speech and language processing for human-machine communications: proceedings of CSI 2015*:195–207
- [21] Hama Saeed M (2023) Improved speech emotion classification using deep neural network. *Circuits Syst Signal Proc* 42(12):7357–7376
- [22] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English,” *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0196391.
- [23] L. R. Rabiner and R.W. Schafer, *Introduction to Digital Speech Processing (Foundations Trends Signal Process)*. Hanover, MA, USA: Now Publishers Inc. 2007.
- [24] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.
- [25] GammatoneFilterBank, MathWorks, Natick, MA, USA, 2024.
- [26] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, “An auditory-based feature for robust speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. 4625–4628.
- [27] X. Valero and F. Alias, “Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification,” *IEEE Trans. Multimedia*, vol. 14, no. 6, pp. 1684–1689, Dec. 2012.
- [28] J. Ramirez, J. C. Segura, C. Benitez, Á. de la Torre, and A. Rubio, “Efficient voice activity detection algorithms using long-term speech information,” *Speech Commun.*, vol. 42, nos. 3–4, pp. 271–287, Apr. 2004.
- [29] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, “Language identification: A tutorial,” *IEEE Circuits Syst. Mag.*, vol. 11, no. 2, pp. 82–108, 2nd Quart., 2011.
- [30] T. Ganchev, *Contemporary Methods for Speech Parameterization*. New York, NY, USA: Springer, 2011.
- [31] A. M. Darling. *Properties Implement. Gammatone Filters: A Tutorial*. Accessed: Mar. 4, 2023. [Online]. Available: <https://www.phon.ucl.ac.uk/home/shl5/Darling1991-GammatoneFilter.pdf>
- [32] R. Islam and M. Tarique, “Investigating the performance of gammatone filters and their applicability to design cochlear implant processing system,” *Designs*, vol. 8, no. 1, pp. 1–21, Feb. 2024.
- [33] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Res.*, vol. 47, nos. 1–2, pp. 103–138, Aug. 1990.
- [34] N. Sugan, N. S. S. Srinivas, N. Kar, L. S. Kumar, M. K. Nath, and A. Kanhe, “Performance comparison of different cepstral features for speech emotion recognition,” in *Proc. Int. CET Conf. Control, Commun., Comput. (IC4)*, Jul. 2018, pp. 266–271.
- [35] A. Géron, *Hands-on Machine Learning With Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA, USA: O’Reilly Media, 2022.
- [36] N. S. S. Srinivas, N. Sugan, L. S. Kumar, M. K. Nath, and A. Kanhe, “Speaker-independent Japanese isolated speech word recognition using TDRC features,” in *Proc. Int. CET Conf. Control, Commun., Comput. (IC4)*, Japan, Jul. 2018, pp. 278–283.