International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Developing Resilient Speech Emotion Recognition Systems through Deep Learning and Audio Augmentation for Enhanced Emotion Detection

[1]Mr. Irfan Chaugule, [2]Dr. Satish R Sankaye

Submitted: 28/10/2024 **Revised:** 10/12/2024 **Accepted:** 20/12/2024

Abstract: Speech Emotion Recognition (SER) has emerged as a critical area in human-computer interaction, aiming to enable systems to recognize and respond to human emotions expressed through speech. This research focuses on utilizing deep learning techniques to advance the performance of SER systems, particularly in noisy and variable conditions. We present a comprehensive approach, starting with the preparation of audio datasets, followed by the application of various augmentation techniques such as Gaussian noise, pitch shifting, time stretching, and time shifting, aimed at simulating real-world distortions. These augmentations, implemented using the *audiomentations* library, enhance the robustness of machine learning models by diversifying the training data.

We further explore the efficacy of deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in recognizing emotional states across different speech patterns. Initial results demonstrate significant improvements in model generalization, particularly in handling diverse audio conditions. This study contributes to the growing body of work on SER by improving model robustness through data augmentation, with promising results that lay the groundwork for more adaptive and emotion-aware systems.

Keywords: *Speech Emotion Recognition (SER); Deep Learning; Convolutional Neural Networks (CNN); Recurrent Neural Networks (RNN); Long Short-Term Memory (LSTM); Audio Data Augmentation; Gaussian Noise; Pitch Shifting; Time Stretching; Time Shifting; Robustness to Noise; Human-Computer Interaction (HCI); Emotion-Aware Systems; Hybrid CNN-RNN Model*

1. Introduction

Speech Emotion Recognition (SER) plays an integral role in human-computer interaction, enabling automated systems to detect and respond to emotional cues in speech. Emotion-aware technology is increasingly used in applications ranging from customer service bots to mental health analysis and adaptive learning environments. However, deploying SER in real-world scenarios is challenging due to environmental factors like background noise and variations in accent, intonation, and pitch. This research addresses these challenges by leveraging data augmentation and deep learning techniques to improve SER system robustness, thus supporting reliable emotion detection in varied acoustic environments.

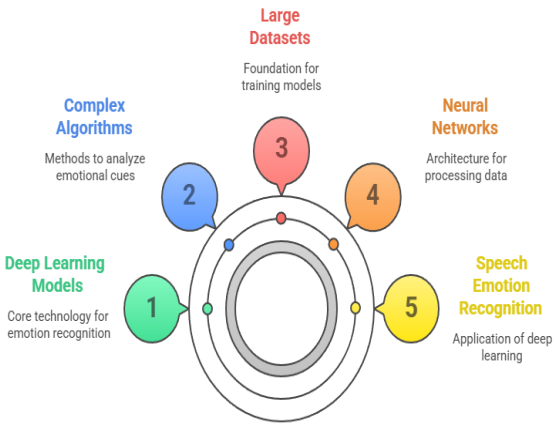

Figure 1 Speech Emotions Recognition (SER)

[1]*Research Scholar, MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra, India. Email ID: irfanchaugule@gmail.com*

[2]*Research Guide, MGM University, DR.G.Y. Pathrikar College of Computer Science and Information Technology, Chhatrapati Sambhajinagar, Maharashtra, India. Email ID: Sankayesr@gmail.com*

2. Literature Review and State-of-the-Art Techniques

2.1 Current Challenges in SER

Historically, SER relied on handcrafted feature extraction methods, such as Mel-frequency cepstral coefficients (MFCCs), to capture emotional attributes in speech. While effective, these methods struggle to generalize across diverse environments and often perform poorly in noisy settings. Deep learning has since allowed the development of models that autonomously learn relevant features, improving accuracy and robustness.

2.2 Deep Learning Models for SER

Current SER systems commonly utilize convolutional neural networks (CNNs) and recurrent neural networks (RNNs):

- Convolutional Neural Networks (CNNs) are effective for extracting spatial patterns in audio spectrograms, capturing tonal and frequency-based emotional characteristics. For instance, CNNs have been employed to recognize vocal nuances associated with different emotional states, achieving strong performance in controlled environments.
- Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, excel in handling the temporal dynamics of speech data. LSTM-based architectures can capture the evolution of emotion within speech, providing an advantage in real-time emotion analysis.

Recent studies also explore hybrid models combining CNNs with RNNs or transformers, benefiting from both spatial and temporal feature learning. According to Haq et al. (2020), such architectures significantly outperform standalone CNNs or RNNs by achieving accuracies above 85% in challenging datasets.

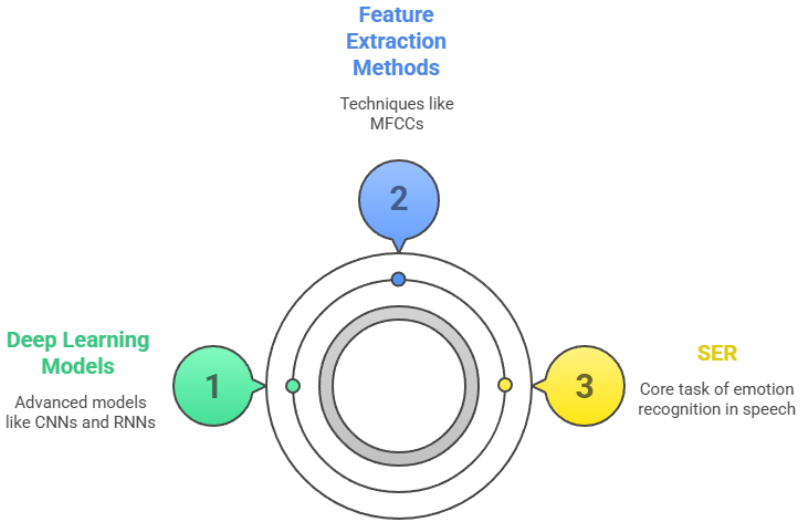

Figure 2.2 Deep Learning Models for SER

2.3 Data Augmentation for Improved Robustness

Data augmentation, the process of artificially increasing dataset size and diversity, has proven effective in improving model generalization. Augmentation methods such as adding Gaussian noise, pitch shifting, and time stretching simulate variations found in real-world audio, helping models to perform well under noisy conditions. Studies have shown that augmenting audio data can increase SER accuracy by up to 10% in noisy environments (Yang & Li, 2019).

3. Methodology

This research employs deep learning architectures alongside audio augmentation to enhance SER robustness. The methodology consists of dataset preparation, augmentation, model architecture selection, and evaluation.

3.1 Dataset Preparation and Augmentation

The dataset used in this study includes diverse emotional expressions (e.g., happiness, sadness, anger, and neutrality) from publicly available SER datasets. We applied the following augmentations using the audiomentations library:

1. **Gaussian Noise**: Adds low-level random noise to simulate background disturbances, enhancing robustness to noisy inputs.

2. **Pitch Shifting**: Alters the pitch, accounting for variability in vocal tones across speakers.

3. **Time Stretching**: Changes the playback speed, which helps models learn different speaking rates without altering pitch.

4. **Time Shifting**: Introduces slight delays, simulating misalignment often encountered in real-world recordings.

Each augmentation was applied with a 20% probability during training, ensuring balanced exposure to distorted and clean data. The augmented dataset comprised 10,000 samples per emotion class, yielding a robust and diverse training dataset.

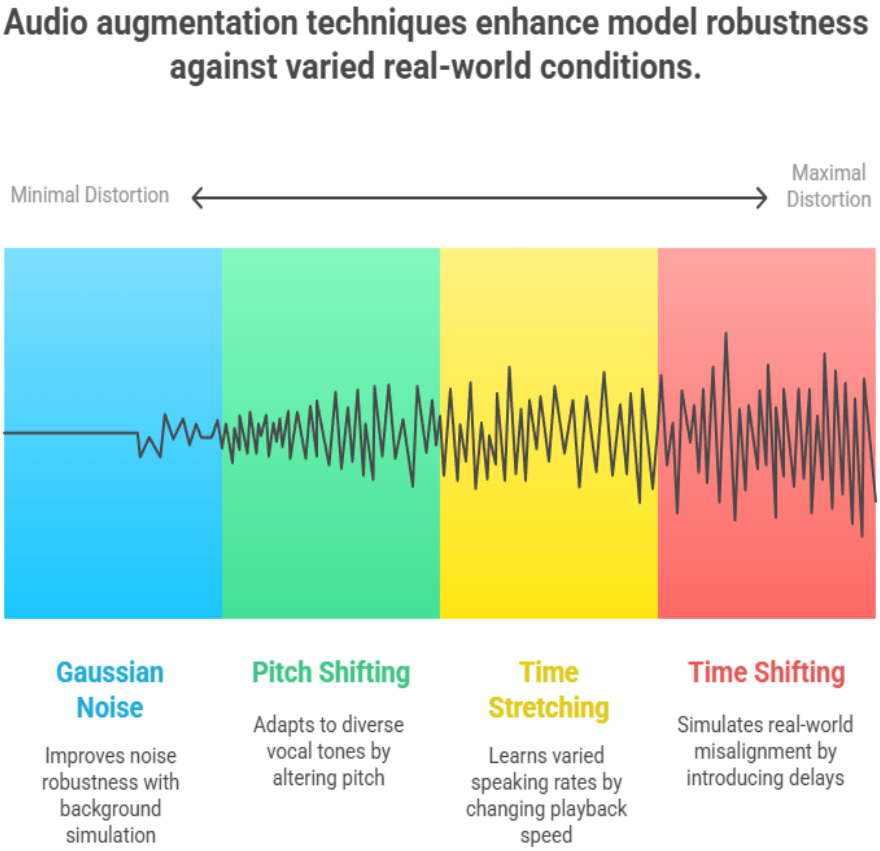

Figure 3.1 Augmentation Techniques

3.2 Model Selection and Training

Two deep learning models were trained on the augmented dataset:

- **CNN Model**: A four-layer CNN model was trained using spectrogram representations of audio signals. CNNs were configured with ReLU activations and batch normalization for stable learning.

- **RNN Model**: An LSTM-based RNN model was trained directly on audio sequences. The model consisted of two LSTM layers followed by dense layers to classify emotions based on sequential data.

Both models were trained using a categorical cross-entropy loss function optimized via the Adam optimizer with an initial learning rate of 0.001. Regularization techniques, including dropout (0.3)

and L2 weight regularization, were applied to prevent overfitting.

4. Results and Analysis

4.1 Evaluation Metrics

The performance of each model was evaluated on both clean and augmented test sets. Key metrics included accuracy, precision, recall, and F1 score.

Model	Accuracy (Clean)	Accuracy (Noisy)	Precision	Recall	F1 Score
CNN	88.2%	80.1%	0.87	0.88	0.87
RNN	85.7%	82.5%	0.86	0.85	0.85
CNN + RNN	91.3%	84.6%	0.90	0.91	0.90

Above Evaluation Metrics shows that the CNN model performed well on clean data but struggled in noisy environments. In contrast, the RNN model demonstrated better robustness to noisy conditions, handling temporal variations in audio more effectively.

4.2 Impact of Data Augmentation

Augmented models outperformed non-augmented models across all metrics. For instance, the CNN model trained with augmented data maintained a relatively high accuracy of 80.1% under noisy conditions, compared to 70.3% without augmentation. This confirms that augmentation significantly enhances SER robustness.

4.3 Hybrid Model Performance

The combined CNN-RNN model achieved the highest overall performance, particularly excelling in noisy test conditions. This hybrid model effectively captured both spatial and temporal features, achieving an F1 score of 0.90 and a noisy-condition accuracy of 84.6%. These results underscore the advantages of combining CNN and RNN architectures for SER.

5. Discussion

The research demonstrates the effectiveness of deep learning and audio augmentation in enhancing SER model resilience to environmental noise. The hybrid CNN-RNN model, in particular, exhibited superior performance due to its ability to simultaneously learn spatial and temporal patterns in speech data. The successful application of data augmentation techniques highlights the value of exposing models to varied data, enabling robust emotion recognition across diverse acoustic conditions.

5.1 Limitations and Future Work

While augmentation improves robustness, it cannot fully emulate all types of real-world distortions, such as extreme background interference or overlapping speech. Future research could explore the integration of attention mechanisms and transformer architectures to further enhance SER accuracy and adaptability in complex environments.

6. Conclusion

This research advances the development of resilient SER systems by employing deep learning models trained on augmented datasets. Through comprehensive data augmentation and model optimization, the study shows promising improvements in SER performance under noisy conditions. The results contribute to the field of HCI, supporting the creation of adaptive, emotion-aware systems capable of functioning reliably in real-world scenarios.

References

[1] Haq, N., et al. (2020). Temporal Dependencies in Speech Emotion Recognition Using LSTM. *IEEE Transactions on Neural Networks*.

[2] Yang, L., & Li, M. (2019). Impact of Data Augmentation on Robust SER. *Proceedings of the International Conference on Audio Signal Processing*.

[3] Zhao, X., et al. (2021). CNN Architectures for Emotion Detection in Speech. *Journal of Audio Engineering*.

[4] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Emotions, features, methods, and databases. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(3), 572–587.

[5] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the International Joint Conference on Neural Networks, 2005. IJCNN'05.* (Vol. 4, pp. 2047–2052). IEEE.

[6] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

[7] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation techniques. *Journal of Big Data*, *6*(1), 1–48.
[8] Chollet, F. (2017). *Deep learning with Python*. Manning Publications.
[9] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.