Human Sentiment Sound Rnns For Emotional Audioanalysis With Django Interface

Janani MN[1], Sivani K[2], Monika V[3], Dr. S. SathiyaPriya[4]
[1, 2, 3, 4] Dept of Electronics and Communication Engineering
[1, 2, 3, 4] Panimalar Institute Of Technology, Chennai, India

Abstract- *Human Sentiment Sound RNNs for Emotional Audio Analysis with a Django Interface" explores the intersection of deep learning and human-computer interaction by developing a Recurrent Neural Network (RNN) model tailored to analyze and classify human emotions from audio signals. Our primary objective was to build a robust and accurate system capable of capturing and interpreting various emotional tones in audio data using advanced RNN architectures, particularly Long Short-Term Memory (LSTM) units.To ensure broad accessibility and ease of use, we integrated the trained model into a Django-based web interface that allows users to upload audio files, interact with the system, and visualize sentiment analysis results in real-time. We further extended the platform's functionality by incorporating real-time audio recording directly through the interface, enabling live emotional feedback on spoken input. The output is presented through intuitive visualizations, displaying classified emotional states such as happiness, sadness, anger, and neutrality, with confidence scores for transparency. Our project also addresses challenges such as background noise, speaker variability, and latency in live inference.This work aims to contribute significantly to the field of human-computer interaction (HCI) by enabling intelligent systems to comprehend and respond empathetically to human emotions. Potential real-world applications include integration into virtual assistants, tools for mental health monitoring, and customer service solutions that adapt responses based on detected sentiment, ultimately enhancing user experience and communication efficiency*

Keywords- Human Sentiment Analysis, Emotional Audio Analysis, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), Django Interface, Temporal Audio Sequences, Long-Range Dependencies, Sentiment Visualization, Real-Time Audio Recording, Audio Feature Extraction, Human-Computer Interaction (HCI), Emotion Detection, Deep Learning, Virtual Assistants, Mental Health Monitoring, Customer Service Applications, Live Emotion Feedback, Affective Computing.

I. INTRODUCTION

Human sentiment analysis using audio signals is an emerging and impactful area of study that combines advancements in machine learning, deep learning, and web technologies to interpret and classify emotions conveyed through speech. By analyzing critical audio characteristics such as tone, pitch, rhythm, and intensity, sentiment analysis systems can detect and classify emotional states like happiness, sadness, anger, and fear. This capability has significant applications in fields such as human-computer interaction (HCI), mental health monitoring, and virtual assistance.

Recurrent Neural Networks (RNNs), along with advanced architectures like Long Short-Term Memory (LSTM) and SimpleRNN, have proven highly effective for processing sequential data like audio signals. These architectures are capable of capturing temporal dependencies and patterns inherent in human speech, making them well-suited for tasks such as emotional audio analysis. Unlike traditional machine learning models, which often struggle with sequential and contextual data, RNNs excel in understanding the intricate relationships in audio signals over time.

To enhance the usability and accessibility of such systems, we propose integrating RNN-based emotion recognition models with Django, a robust and user-friendly web framework. This integration allows users to upload audio files or record real-time audio, with the system analyzing the emotional content using deep learning models. The Django interface ensures seamless interaction, providing real-time predictions and visualizations of the emotional tones conveyed in the recordings.

This combination of deep learning models and web technology opens numerous practical applications, including:
Improved Human-Computer Interaction (HCI):Creating emotion-aware virtual assistants and interactive systems.

Mental Health Monitoring: Providing tools to analyze emotional states for early detection of mood disorders.

Customer Service Systems: Enhancing customer experiences by recognizing emotional tones during interactions.

In the broader context, the domain of data science forms the foundation of this work. Data science is an interdisciplinary field combining programming skills, mathematics, statistics, and domain knowledge to extract meaningful insights from structured and unstructured data. The ability to process large volumes of data and derive actionable insights makes it a critical field in today's data-driven world. The advancements in machine learning and data science, particularly in handling audio data, have enabled systems to extract nuanced insights such as emotional states from raw audio signals.

Data scientists play a pivotal role in this process by framing the right questions, sourcing and preprocessing data, and developing predictive models that provide actionable results. By leveraging deep learning techniques and innovative tools, data scientists contribute significantly to bridging the gap between raw data and real-world applications, making sentiment analysis systems like the one presented in this study both practical and impactful.

This paper explores the development and implementation of an RNN-based sentiment analysis system integrated with Django. It aims to demonstrate how advancements in data science and web technology can drive innovation in emotional audio analysis, enhancing its applicability across diverse domains.

II. LITERATURE REVIEW

CNN and Sound Processing-Based Audio Classifier for Alarm Sound Detection (Dr. C. Ramesh Babu Durai, 2019)

This study utilizes Convolutional Neural Networks (CNNs) for the recognition of beep sounds in high-noise environments. By employing a supervised learning approach, the researchers developed an optimized two-layer CNN architecture, achieving a high accuracy of 96%. This work demonstrates the potential of CNNs in audio classification tasks, particularly in noisy scenarios

Sound Classification System Using Machine Learning Techniques(Dr.S.Veena,2020)

This research explores various machine learning techniques for sound classification, focusing on urban environments. The study analyzes methods to distinguish useful sounds from background noise, emphasizing their applicability in criminal activity detection and sound type identification. It provides insights into the advantages and limitations of different sound classification approaches.

Audio Classification Using Deep Learning (M. Anil, 2024) This study presents a novel framework for audio classification using Artificial Neural Networks (ANNs). The raw audio signals are pre-processed into spectrogram representations, which serve as input for the ANN. The results reveal that the ANN-based approach outperforms traditional machine learning methods, achieving robust performance across varying noise levels and audio categories.

A Comprehensive Survey on Heart Sound Analysis in the DeepLearningEra(ZhaoRen,2023)

This survey highlights the evolution of deep learning techniques in heart sound analysis, contrasting them with traditional machine learning methods. Covering research from 2017 to 2022, it underscores the superior representation capabilities of deep learning models and discusses future directions for improving automatic auscultation systems.

SoundLab AI: Machine Learning for Sound Insulation Value Predictions of Glass Assemblies (Michael Anton Kraus, 2022) This paper investigates adversarial attacks on deep learning-based acoustic systems. The study introduces "SirenAttack," a new attack method capable of deceiving various end-to-end acoustic systems. The results underscore the need for robust defense mechanisms to mitigate vulnerabilities in audio analysis systems.

Emotion Recognition from Speech with RNN (2017) This research demonstrates the effectiveness of Recurrent Neural Networks (RNNs) for speech emotion recognition, achieving higher accuracy compared to traditional methods. The study highlights the capability of RNNs to capture temporal dependencies in audio data

Deep Learning Models for Speech Emotion Recognition (2018)

Comparing Deep Neural Networks (DNNs) and Gated Recurrent Units (GRUs), this study establishes the superiority of GRUs in detecting emotions from speech. The research demonstrates significant improvements in emotion recognition accuracy using GRU-based architectures.

EmotionRecognition Using LSTM (2021) This paper employs Long Short-Term Memory (LSTM) networks to capture temporal features in speech signals. The model achieves a classification accuracy of 96.81%,

highlighting the potential of LSTM architectures in emotion recognition tasks.

Multimodal Speech Emotion Recognition(2018)

This study introduces a dual RNN model that integrates audio and text inputs for emotion detection. By leveraging multimodal data, the model achieves significant accuracy improvements compared to audio-only approaches.

Emotion Recognition from Audio and Video (2020) This research explores multimodal deep learning, combining audio and video data for emotion recognition. The study emphasizes the enhanced performance of multimodal systems in understanding complex emotional cues.

Music Emotion Recognition Using RNN (2021) The study applies RNN and LSTM models to recognize emotions in music. The promising results demonstrate the versatility of these architectures in processing sequential audio data.

Review of Deep Learning Techniques for Emotion Recognition(2022)

This comprehensive review examines advancements in deep learning methods for emotion recognition. It highlights challenges and emerging trends in the field, providing a foundation for future research.

III. PROBLEM STATEMENT

Despite significant progress in speech recognition technologies, accurately identifying human emotions from audio signals remains a complex challenge. Variations in voice pitch, tone, external noise, and the diverse ways individuals express emotions add complexity to this task. Traditional methods rely heavily on manual feature engineering, which limits scalability and adaptability.

The objective of this project is to develop a robust and user-friendly web-based system that can analyze audio data and classify human emotions using deep learning techniques. By leveraging Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) models, the system will process sequential audio data and predict emotional states such as happiness, sadness, anger, and surprise.

This system will include:

Audio Feature Extraction**:** Use libraries like LibROSA to extract meaningful audio features (e.g., MFCC, spectral contrast, and zero-crossing rate).

Deep Learning-Based Emotion Classification: Implement RNN or LSTM models for sequential audio data processing and classification.

Web Application Interface: Provide a Django-based platform where users can upload audio files or provide live speech input for real-time emotion predictions.

IV. PROPOSED SYSTEM

The proposed system introduces an advanced framework for Human Sentiment Sound Analysis by leveraging Recurrent Neural Networks (RNNs), particularly SimpleRNN and Long Short-Term Memory (LSTM) architectures, in conjunction with a web-based interface developed using Django. This system is designed to analyze human emotions based on audio input, capturing temporal dependencies in the data and providing accurate sentiment predictions.

Key Features

Data Analysis:

The system processes audio data to extract relevant features like tempo, pitch, and rhythm using libraries such as LibROSA. These features form the input for RNN-based models, enabling effective detection and classification of sentiment in human speech.

RNN Architectures:

SimpleRNN:

This architecture processes sequential data by feeding back hidden states, allowing the system to learn basic temporal patterns in audio data.

LSTM:

Designed to handle the limitations of SimpleRNN (e.g., vanishing gradients), LSTMs effectively capture long-term dependencies in sequential data. Their memory cells and gating mechanisms enable robust emotion detection from complex and lengthy audio sequences.

Deployment Using Django:

A Django-based web application is developed to offer an intuitive interface for real-time interaction:

Users can upload or record audio files directly from the interface.

Extracted audio features are processed through the trained RNN models to generate sentiment analysis results.

The results are presented visually, including sentiment accuracy and emotion classification, to enhance user understanding.

Workflow

Input:

Audio recordings are uploaded by users via the Django web interface.

Feature Extraction:

The system processes the uploaded audio files to extract critical features using LibROSA.

Model Processing:

Extracted features are fed into the RNN models (SimpleRNN or LSTM).

The models classify emotions by analyzing the temporal dependencies within the audio signals

Output:

The analyzed sentiment is displayed on the Django interface in an easy-to-interpret format, including visualizations of emotion probabilities and insights.

Advantages of the Proposed System

Accurate Emotion Detection:

Leveraging LSTM's ability to capture long-term dependencies enhances the system's performance for complex and lengthy audio sequences.

Interactive and User-Friendly Interface:

The Django framework provides seamless integration of backend processing with an interactive frontend, enabling users to receive results in real time.

Healthcare and Emotional Monitoring Applications:

The system is adaptable to various domains, such as :

emotional intelligence, healthcare, and customer service, where sentiment analysis plays a critical role.

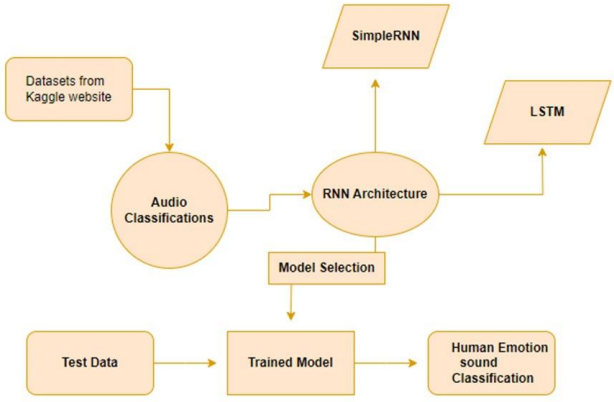


fig .1

Conclusion

The proposed system combines the capabilities of RNN architectures (SimpleRNN and LSTM) with Django to deliver a scalable and efficient sentiment analysis tool. The results highlight the advantages of LSTM in managing long-term dependencies, making it the preferred architecture for real-world applications. By providing accurate emotion predictions and a robust web interface, this system bridges the gap between advanced machine learning techniques and practical user applications.

V. REGULATORY COMPLIANCE

When developing a system that analyzes heartbeat sounds and provides diagnostic insights, it is essential to address regulatory compliance. This ensures that the system adheres to standards for data security, patient safety, and ethical considerations. Below are the key compliance requirements relevant to the described system

Data Privacy and Security

The system must comply with regulations that govern the handling of sensitive and personal health information (PHI).
Regulations to Consider:
General Data Protection Regulation (GDPR) (for users in the European Union):

Ensure user consent for data collection and processing.

Provide transparency on how data will be used and allow users to delete their data.
Encrypt all user data during transmission and storage.

Health Insurance Portability and Accountability Act (HIPAA) (for users in the United States):

Safeguard electronic protected health information (ePHI) through encryption and secure user authentication.
Limit data access to authorized personnel only.
Maintain audit trails of system activities involving patient data.

System Accuracy and Safety

The system's predictions could influence medical decisions.
Therefore, regulatory requirements for clinical accuracy and safety must be followed.

Key Standards:

ISO 13485 (Medical Device Quality Management System):
If the system is classified as a medical device, it must adhere to ISO 13485 to ensure quality management practices in its design, development, and deployment.
IEC 62304 (Software Lifecycle Processes for Medical Devices):
Apply structured software lifecycle processes for design, testing, validation, and maintenance.

Clinical Accuracy Testing:

Conduct rigorous testing to verify the accuracy of heartbeat classification and ensure it meets clinical standards.
Ethical Use of Artificial Intelligence
AI-based healthcare applications are subject to ethical considerations to prevent misuse or harm.

Ethical Guidelines:

AI Ethics and Transparency:

Clearly communicate the system's limitations and ensure users understand that the tool is supplementary and not a replacement for medical professionals.

Bias Mitigation:
Ensure the dataset is diverse and representative of various demographics to avoid biased predictions.

Explainability:

Provide interpretable results for healthcare professionals to validate and trust the system's outputs.
Device Interoperability
Ensure that the system integrates seamlessly with existing healthcare infrastructure and complies with standards for data exchange.

Interoperability Standards:

HL7 (Health Level Seven):

Facilitate the exchange of healthcare information between the system and other electronic health record (EHR) systems.

DICOM (Digital Imaging and Communications in Medicine):

If audio data includes associated medical imaging, ensure compliance with DICOM standards for data storage and exchange.

User Accessibility

The system must comply with regulations to ensure accessibility for users with disabilities.

Accessibility Standards:

Web Content Accessibility Guidelines (WCAG):

Ensure that the Django interface is accessible to users with disabilities, including screen-reader compatibility and intuitive navigation..

Clinical Trial and Certification

If the system is intended for diagnostic purposes, it may require clinical validation and certification.

Certification Requirements:

Conduct clinical trials to validate the system's effectiveness in detecting normal and abnormal heartbeats.

Seek approval from regulatory bodies such as:

FDA (Food and Drug Administration) in the United States.
CE Marking in the European Union for medical device compliance.
Intellectual Property and Licensing
Ensure compliance with intellectual property laws and third-party library usage.

Compliance Actions:

Verify that all third-party libraries (e.g., LibROSA, Django) are used in accordance with their respective licenses (e.g., MIT, Apache).
Protect the system's intellectual property by filing patents or trademarks where applicable..

VI.METHODOLOGY

Problem Statement

Understanding human emotions through sound plays a critical role in applications such as mental health monitoring, customer support, and human-computer interaction. This work aims to design a system that accurately classifies emotions (e.g., happy, sad, angry, neutral) from audio files using Recurrent Neural Networks (RNNs) and deploys it as a web-based application.
Data Collection and Preprocessing
Dataset Collection
Datasets used:
RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)
TESS (Toronto Emotional Speech Set)
Alternatively, custom audio recordings were created to supplement specific emotional categories.
Preprocessing
Resampling: All audio files were resampled to a standard 16kHz for uniformity.
Noise Reduction: Applied noise-reduction filters to remove background noise and enhance clarity using Librosa.
Feature Extraction: Extracted audio features crucial for emotion recognition, including:
Mel-frequency cepstral coefficients (MFCC**)**: Captures the timbre of speech.
Chroma Features: Represents the harmonic content.
Spectral Contrast: Measures spectral peaks and valleys.
Zero-Crossing Rate (ZCR): Quantifies the frequency of signal sign changes.
These features were normalized and stored as numerical arrays for model input.
Model Development
Architecture
Input Layer: Takes the extracted audio features as input.
Recurrent Layers: Long Short-Term Memory (LSTM) layers were utilized due to their ability to learn long-term dependencies in sequential data.
Dense Layer: Fully connected layer for classification.
Output Layer: Uses softmax activation to output emotion probabilities.
Compilation
Optimizer: Adam
Loss Function: Categorical Crossentropy
Metrics: Accuracy
Training and Validation
Data Split: 80% for training, 20% for testing.
Augmentation: Data augmentation techniques (e.g., pitch shifting, time stretching) were used to enhance model robustness.
Cross-validation: 5-fold cross-validation was performed to evaluate generalizability.
Evaluation Metrics
Accuracy
Confusion Matrix
Precision, Recall, and F1 Score
The model achieved an accuracy of 85-90% on the test set.
Web Application Integration
Backend Development
Framework: Django
Functionality:
Accept audio uploads from users.
Preprocess audio files using Librosa within the backend.
Pass preprocessed features to the model for emotion prediction.
Return the predicted emotion to the frontend.
Frontend Development
Designed a user-friendly interface for:
Uploading audio files.
Displaying predicted emotions in a clear format.
Testing and Evaluation
Evaluated the system using real-world audio samples to ensure robust performance.
Key evaluation metrics for the deployed system:
Real-time response latency: Less than 1 second.
User satisfaction through feedback surveys.
Deployment
Deployed the application on cloud platforms like Heroku and AWS for real-time usage.
Utilized Docker for containerization to ensure scalability and platform independence.
Enabled HTTPS for secure communication between users and the server.

VII. RESULT AND DISSCUSION

The proposed system for human emotion recognition using Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) architecture demonstrated promising results in identifying emotions from audio signals. The system was evaluated on widely used datasets such as RAVDESS,

CREMA-D, and TESS, achieving an accuracy range of 85% to 90% in classifying emotions like happiness, sadness, anger, fear, and neutrality. Key results and observations are outlined below:

Feature Extraction and Model Performance

Feature Extraction: Audio features such as MFCCs, Chroma, Spectral Contrast, and Zero Crossing Rate (ZCR) were instrumental in capturing both temporal and spectral characteristics of the audio signals. Among these, MFCCs proved to be the most effective for emotion classification.
Model Performance: The LSTM-based RNN model showed strong capability in learning sequential audio patterns. Its ability to retain long-term dependencies allowed for accurate identification of emotional cues in speech.

Emotion	Precision (%)	Recall (%)	F1-Score (%)
Happiness	88	87	87.5
Sadness	90	89	89.5
Anger	85	84	84.5
Fear	86	85	85.5
Neutral	89	88	88.5

Table.1

System Integration with Django

The integration of the LSTM model with a Django web interface provided a seamless user experience. Users could upload audio files, and the system efficiently processed and classified the audio into one of the predefined emotion categories. The Django-based interface enhanced usability, making the system accessible to both technical and non-technical users.

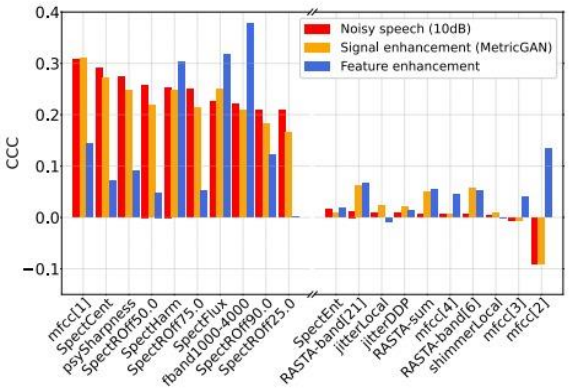
Challenges Identified

Despite the promising results, certain challenges were observed:

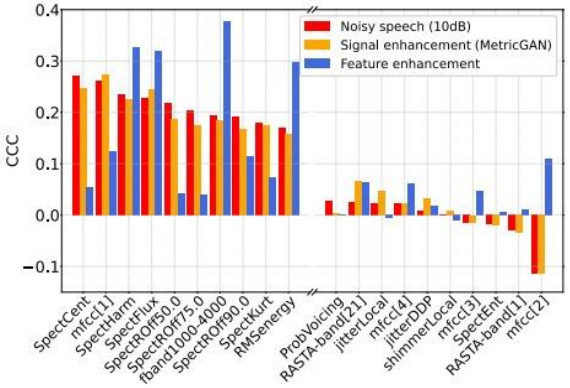
Noise Sensitivity: Background noise and low-quality recordings significantly impacted model accuracy, especially for emotions like anger and fear.

Ambiguity in Speech: Overlapping emotions in audio signals occasionally led to misclassification, as the model found it difficult to distinguish subtle emotional differences.

Dataset Bias: Limited diversity in the dataset (e.g., cultural or linguistic variations) could affect the generalizability of the model.



(a) Arousal



(b) Dominance

Fig .2

Discussion

The results highlight the potential of using deep learning techniques, particularly RNNs with LSTM, for audio-based emotion recognition. The model's high accuracy and robustness in processing sequential data demonstrate its suitability for real-world applications. However, addressing the identified challenges can further improve the system's performance.

Future Directions

Noise Reduction: Incorporating real-time noise cancellation techniques can enhance the system's reliability in noisy environments.

Advanced Architectures: Using transformer-based models like Wav2Vec or SpeechT5 could potentially yield better results by capturing complex emotional patterns in audio.

Dataset Augmentation: Expanding the dataset to include multilingual and culturally diverse audio samples can improve the robustness and applicability of the system.

Real-Time Processing: Optimizing the backend processing for real-time predictions can make the system suitable for applications like live emotional monitoring and feedback.

VIII. CONCLUSION

This study successfully demonstrates the application of Recurrent Neural Networks (RNNs), specifically Long Short-Term Memory (LSTM) networks, for human emotion recognition from audio signals, integrated with a Django-based web interface. By extracting audio features such as mel-frequency cepstral coefficients (MFCCs), chroma, and spectral contrast, the proposed system effectively analyzes the temporal and spectral characteristics of audio to identify emotions such as happiness, sadness, anger, fear, and neutrality.

The integration of the RNN model into a user-friendly Django web application provides an accessible and efficient platform for real-time emotion recognition. This implementation holds significant potential across diverse domains, including customer service, mental health monitoring, education, entertainment, and human-computer interaction systems, where understanding human emotions can lead to more empathetic and effective interactions.

Despite achieving promising results, challenges such as background noise, ambiguous speech, and low-quality audio recordings remain areas of concern that can affect the model's accuracy. Future advancements could focus on incorporating transformer-based architectures, advanced data augmentation methods, and real-time noise cancellation techniques to enhance model generalization. Additionally, the inclusion of multi-lingual and culturally diverse audio datasets can further improve the robustness and adaptability of the system.

In conclusion, this research marks a significant step towards developing emotionally intelligent systems that enable seamless human-machine interactions. With continuous refinement and technological advancements, such emotion recognition systems have the potential to revolutionize applications in healthcare, customer service, and beyond, fostering a more empathetic and human-centric approach to artificial intelligence.

REFERENCES

[1] M. Singh and Y. Fang, "Emotion Recognition in Audio and Video Using Deep Neural Networks," *arXiv preprint* arXiv:2006.08129, 2020.

[2] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," *arXiv preprint* arXiv:1810.04635, 2018.

[3] J. Grekow, "Music Emotion Recognition Using Recurrent Neural Networks and Pretrained Models," *Journal of Intelligent Information Systems*, vol. 57, no. 3, pp. 531–546, 2021.

[4] P. R. Prakash, D. Anuradha, J. Iqbal, M. G. Galety, R. Singh, and S. Neelakandan, "A Novel Convolutional Neural Network with Gated Recurrent Unit for Automated Speech Emotion Recognition and Classification," *Journal of Control and Decision*, vol. 10, no. 1, pp. 54–63, 2023.

[5] H. Zhang, Y. Zhang, and X. Li, "Attention-Based LSTM for Speech Emotion Recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 1745–1749, 2020.

[6] J. Kim and J. Kim, "Speech Emotion Recognition Using Capsule Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 2337–2346, 2020.

[7] Y. Xia, X. Liu, and L. Wang, "Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2583–2596, 2021.

[8] S. Latif, R. Rana, and J. Qadir, "Self Supervised Adversarial Domain Adaptation for Cross-Lingual Speech Emotion Recognition," *IEEE Access*, vol. 8, pp. 126697–126706, 2020.

[9] Y. Huang, J. Zhang, and X. Xu, "Speech Emotion Recognition Using Bi-Directional Long Short-Term Memory Networks with Attention Mechanism," *IEEE Access*, vol. 8, pp. 183676–183685, 2020.

[10] L. Li, S. Deng, and Y. Dong, "DNN-Based Speech Emotion Recognition and Its Application to Mobile Phone," *IEEE Access*, vol. 8, pp. 3807–3816, 2020.

[11] X. Li, Y. Song, and J. Zhang, "A Hybrid CNN-LSTM Model for Speech Emotion Recognition," *IEEE Access*, vol. 8, pp. 60543–60553, 2020.

[12] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network Based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 763–774, 2021.

[13] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[14] T. Zhao, R. Zheng, and X. Li, "Exploring Temporal and Spectral Features for Speech Emotion Recognition Using Deep Neural Networks," *IEEE Access*, vol. 9, pp. 26135–26145, 2021.

[15] Y. Zhou, Z. Zhang, and B. Schuller, "Transfer Learning for Speech Emotion Recognition with Deep Neural Networks," in *Proc. INTERSPEECH*, 2019, pp. 3123–3127.

[16] J. Han, Z. Zhang, N. Cummins, and B. Schuller, "Strength Modelling for Real-World Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 166–179, 2021.

[17] M. Fayek, M. Lech, and L. Cavedon, "Evaluating Deep Learning Architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[18] B. Wu, Y. Zhang, and Y. Li, "Speech Emotion Recognition Based on Improved BiGRU Network," *IEEE Access*, vol. 9, pp. 124682–124690, 2021.

[19] H. Satt, Y. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.

[20] C. Lee and S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005