Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPEECH EMOTION RECOGNITION USING DEEP LEARNING TECHNIQUES

ZABER AL HASSAN AYON, MUHAMMAD ALIFF AHMAD ZAINUDIN, NUR HAFIEZA ISMAIL*, NUR SHAZWANI KAMARUDIN, MUHAMMAD ARY MURTI[2]

Faculty of Computing, University Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia

[2]Electrical Engineering Study Program, School of Electrical Engineering, Telkom University, Indonesia

Corresponding Author: hafieza@umpsa.edu.my*

ABSTRACT

Accurate recognition and interpretation of emotions from speech signals represent a critical frontier in human-computer interaction and affective computing. This research presents a novel Deep Learning (DL) framework for speech emotion recognition that achieves state-of-the-art performance by integrating advanced acoustic feature extraction with sophisticated sequential modeling. The proposed approach leverages Mel-frequency Cepstral Coefficients (MFCC) for robust feature extraction and employs Long Short-Term Memory (LSTM) networks to effectively capture the temporal dynamics inherent in emotional speech patterns. Through rigorous experimentation on the Toronto Emotional Speech Set (TESS), our model demonstrates exceptional accuracy of 98.21%, significantly outperforming traditional Machine Learning (ML) approaches. Comparative analysis reveals the LSTM-based model has superior ability to differentiate between acoustically similar emotions, a persistent challenge in speech emotion recognition systems. The architecture's computational efficiency and robustness to acoustic variability make it particularly well-suited for real-time applications across diverse domains including healthcare monitoring, customer experience management, and intelligent human-machine interfaces. This research advances the field of affective computing by establishing a comprehensive framework that combines acoustic feature engineering with deep sequential learning, offering a reliable solution for emotion recognition from speech signals and laying the groundwork for more intuitive and emotionally intelligent computing systems.

Keywords: *Deep learning, Speech signal, Emotion recognition, LSTM, MFCC.*

1. INTRODUCTION

Speech emotion recognition (SER) has emerged as a pivotal technology within artificial intelligence (AI), particularly in the domain of ML and DL. The ability to detect human emotions through speech offers transformative potential across multiple sectors, including healthcare, customer service, education, and entertainment. By enabling machines to interpret emotional cues from vocal signals, SER systems can enhance human-computer interactions, offering more personalized and empathetic responses. This study aims to explore the development of SER systems through the integration of ML and DL models, examining their capacity to accurately detect emotional states from speech.

Emotions play a fundamental role in human communication, with speech serving as a primary medium for conveying emotional cues. Detecting emotions from speech involves analyzing various acoustic and linguistic features, such as pitch, tone, and prosody, to infer the speaker's emotional state. With the rise of intelligent systems, including virtual assistants, autonomous vehicles, and therapeutic robots, SER has become an essential tool for creating systems capable of responding to users' emotional needs. For instance, in customer service, emotion recognition can help mitigate customer frustration by allowing systems to adjust responses based on detected emotional states, while in mental health, SER systems could monitor and flag early signs of emotional distress, enabling timely interventions [1], [2] .

The rise of ML and DL approaches has significantly improved the accuracy and robustness of SER systems. Traditional approaches to emotion detection often relied on rule-based models that struggled with variability in speech patterns and emotions across different languages and cultures.

However, with the application of advanced ML models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), emotion detection systems can now achieve higher levels of accuracy by learning complex patterns in speech data. CNNs, for instance, have demonstrated their capacity to extract deep acoustic features from raw speech signals, while LSTM networks have been effective in capturing the temporal dependencies in emotional expression over time [3], [4], [5].

LSTM networks, a specialized form of RNNs introduced by Hochreiter and Schmidhuber in 1997, represent a crucial advancement for sequential data processing such as speech [6]. Unlike traditional RNNs, LSTMs effectively address the vanishing gradient problem through an innovative cell architecture that incorporates three essential gating mechanisms: forget gates, input gates, and output gates [6], [7]. This design allows LSTMs to selectively retain relevant historical information while discarding irrelevant details, making them particularly well-suited for modeling the temporal dynamics of emotional speech patterns. The forget gate determines which information from the previous cell state should be discarded, the input gate regulates the flow of new information into the cell state, and the output gate controls what information should be propagated to the output.

In the context of speech emotion recognition, LSTM networks offer distinct advantages due to their ability to maintain an internal memory of relevant emotional cues across extended sequences. This capability enables LSTMs to capture subtle variations in prosody, rhythm, and energy that serve as critical indicators of emotional states [8], [9]. Recent advancements, including bidirectional LSTMs and attention-augmented LSTMs, have further enhanced their effectiveness by enabling models to consider both past and future context when making predictions and focusing on the most emotionally salient portions of speech signals [10], [11]. These enhancements have contributed significantly to improvements in recognition accuracy, particularly for emotions with subtle acoustic differences.

Despite these advancements, challenges remain in the development of SER systems, particularly in terms of dataset diversity and generalizability. Most SER systems are trained on datasets that primarily feature English-speaking individuals, which limits their applicability to other languages and cultural contexts. Furthermore, the scarcity of annotated emotional speech data in non-English languages continues to hinder the development of robust cross-linguistic emotion detection models. Ensuring that SER systems are both scalable and inclusive requires expanding the availability of multi-lingual datasets and developing models capable of handling diverse accents, dialects, and emotional expressions [12], [13].

In addition to language-related challenges, SER systems must contend with individual differences in emotional expression. Research has shown that emotions are expressed differently based on factors such as personality, gender, and even physical health. ML models trained on large and diverse datasets can help mitigate these variations, allowing for more personalized SER systems that adapt to individual speech patterns. For example, personalized SER models can be fine-tuned to recognize subtle emotional cues that might not be evident in generalized systems [14].

The current research endeavors to further advance the field of SER by addressing the key challenges related to dataset diversity, model scalability, and personalization. By developing a DL-based system capable of detecting emotions across different languages and cultural contexts, this study seeks to improve the generalizability of SER systems while maintaining high accuracy. In doing so, the research aims to contribute to the broader application of emotion detection technology, particularly in fields such as healthcare and mental health diagnostics, where the ability to detect emotional states from speech can facilitate more timely and effective interventions.

2. RELATED WORK

SER has evolved significantly with the advancement of ML and DL techniques. Various approaches focusing on feature extraction, model architecture, and classification strategies have been proposed to address the challenges in recognizing emotions from speech signals.

2.1 Hybrid Approaches in SER

Recent research has demonstrated the effectiveness of hybrid models that combine traditional ML with DL techniques. A significant contribution to this area involves hybrid systems combining Support Vector Machines (SVM) with DL models to enhance SER accuracy [15]. These systems typically employ a three-stage approach: signal pre-processing, feature extraction using MFCC, and classification using a combination of Stacked Autoencoders (SAE) and Deep Belief Networks (DBN). The integration of these

techniques leverages the feature extraction capabilities of DL while enhancing classification performance through SVM, resulting in notable improvements across benchmark datasets including the Berlin and RAVDESS databases [16].

Building upon these foundations, contemporary studies have expanded this approach by exploring hybrid models that integrate CNNs with SVMs [11], [17]. In these models, CNNs extract deep hierarchical features from raw audio signals, capturing subtle emotional characteristics that traditional feature extraction methods might miss. Furthermore, research incorporating temporal attention mechanisms has gained significant traction, demonstrating enhanced capability to highlight critical speech segments while improving robustness under adverse acoustic conditions [11], [18], [19].

The success of DL in speech emotion recognition parallels its advancement in other pattern recognition domains. Recent studies on web page classification highlight the versatility of DL for handling diverse data types. One study applied DL for topic-based classification, achieving strong performance through effective feature design [20]. Another utilized CNNs on word cloud image representations, showing how visualized text can be efficiently processed by deep models [21].

2.2 Wavelet Transforms and Advanced Feature Extraction

Wavelet transform techniques have emerged as powerful tools in SER due to their ability to decompose speech signals into multiple scales, providing a more nuanced representation of emotional content [22], [23]. Research utilizing wavelet coefficients has demonstrated the effectiveness of extracting statistical features such as mean, variance, skewness, and kurtosis, which are fundamental in characterizing emotional states. When combined with Extreme Learning Machines (ELM) for classification, this approach has achieved remarkable performance on established databases like Berlin and eNTERFACE.

ELM-based approaches have gained attention due to their computational efficiency when processing large datasets, making them particularly suitable for real-time applications in emotion recognition [24], [25]. However, these methods sometimes exhibit inconsistent performance across different emotional categories and linguistic contexts. Recent studies have addressed these limitations by incorporating ensemble learning methods and attention-based mechanisms, significantly improving classification consistency [11], [25], [26].

2.3 Deep Belief Networks and Advanced Neural Architectures

DBNs have been extensively explored in the context of SER, particularly when integrated with SVMs to address challenges related to emotional expression variability and environmental noise [23], [27], [28]. This integration combines DBNs' capacity to automatically learn high-level representations of emotional features with the robust classification capabilities of SVMs. While these approaches demonstrate good accuracy, they often require substantial labeled datasets for effective training. A limitation that recent research has sought to mitigate through innovative techniques including data augmentation, transfer learning, and semi-supervised learning methodologies [29], [30].

Contemporary advancements have emphasized the integration of DBNs with LSTM networks to enhance temporal feature learning [8], [13], [31]. This combination allows models to effectively capture both the spatial characteristics of speech signals and their temporal evolution, resulting in significant improvements in emotion recognition accuracy across diverse datasets and real-world applications. The temporal modeling capabilities of LSTM networks are particularly valuable in capturing the dynamic nature of emotional expressions in continuous speech.

2.4 Feature Selection and Fusion Techniques

Feature selection and fusion techniques have proven instrumental in optimizing SER performance [32], [33], [34]. Dimensional reduction methods including Principal Component Analysis (PCA) and evolutionary algorithms such as Genetic Algorithms (GA) have been employed to identify the most relevant features while reducing computational complexity. The strategic combination of MFCC for acoustic feature extraction with SVMs for classification, enhanced by sophisticated fusion techniques, enables SER systems to capture critical emotional indicators with greater precision and efficiency.

Recent studies have expanded upon these methodologies by implementing more advanced feature fusion strategies based on deep learning architectures [35], [36], [37]. These approaches facilitate the integration of acoustic, linguistic, and prosodic features, creating multi-modal representations that capture emotional content across different dimensions of speech. Such

comprehensive feature fusion strategies have demonstrated superior accuracy and robustness compared to single-modality approaches, particularly in challenging real-world environments with varying acoustic conditions.

2.5 Attention Mechanisms and Transformer Architectures

The application of attention mechanisms in SER has transformed the field by enabling models to focus on the most emotionally salient parts of speech signals [38], [39], [40]. Research exploring CNNs integrated with temporal attention mechanisms has demonstrated the potential to capture both spatial and temporal relationships in speech, significantly improving recognition performance under variable conditions. These attention-enhanced systems have achieved exceptional accuracy on standard evaluation datasets including Berlin and Emo-DB, highlighting the effectiveness of selective features focusing on emotion recognition tasks.

Transformer architecture represents the cutting edge in SER research, with recent developments pushing the boundaries of what's possible in emotion recognition [10], [41], [42], [43]. These models excel at simultaneously capturing long-range dependencies and processing data in parallel, addressing limitations of earlier RNN approaches. Transformer-based SER systems have demonstrated state-of-the-art performance across multiple benchmarks, combining computational efficiency with superior recognition capabilities. The self-attention mechanisms at the core of these architectures enable them to model complex interactions between different segments of speech, capturing subtle emotional cues that might be missed by other approaches.

2.6 Cross-Cultural and Multilingual SER

A growing body of research addresses the challenges of developing emotion recognition systems that perform consistently across different cultural and linguistic contexts [44], [45], [46]. These studies highlight the impact of cultural variations in emotional expression and perception, emphasizing the need for diverse training datasets that represent multiple languages and cultural backgrounds. Cross-cultural SER research has employed transfer learning and domain adaptation techniques to bridge gaps between source and target domains, enabling more generalizable emotion recognition systems.

Recent multilingual SER approaches have explored language-independent feature extraction methods and universal emotion representation frameworks that perform robustly across linguistic boundaries [47], [48]. These developments are particularly significant for global applications of emotion recognition technology, where systems must operate effectively across diverse linguistic and cultural environments.

2.7 Multimodal Approaches to Emotion Recognition

Multimodal emotion recognition represents an important frontier in affective computing research, combining speech with other modalities such as facial expressions, physiological signals, and text to improve recognition accuracy [38], [49], [50]. These approaches leverage the complementary nature of different emotional cues, compensating for the limitations of single-modality systems. Research in multimodal fusion strategies has demonstrated significant performance improvements over unimodal approaches, particularly in challenging real-world scenarios with varying environmental conditions.

Advanced DL architectures including cross-modal attention mechanisms and graph neural networks have been employed to model complex relationships between different modalities, capturing both intra-modal and inter-modal dependencies in emotional expressions [47], [51]. These sophisticated fusion techniques enable more comprehensive emotional analysis, approximating the holistic perception capabilities of human observers.

2.8 Real-time SER Systems and Applications

The development of real-time SER systems represents a critical research direction with significant practical implications [7], [52]. These studies focus on optimizing the computational efficiency of emotion recognition algorithms while maintaining high accuracy, enabling deployment on resource-constrained platforms including mobile devices and embedded systems. Techniques such as model quantization, knowledge distillation, and efficient neural architecture search have been employed to reduce the computational requirements of SER systems without significantly compromising performance.

Application-specific SER research has explored tailored solutions for domains including healthcare monitoring, customer service interaction analysis, educational technology, and automotive systems [53], [54]. These specialized applications address domain-specific challenges and

requirements, optimizing emotion recognition for particular use cases and operational contexts.

2.9 Knowledge Gaps and Research Opportunities

Despite substantial advancements in speech emotion recognition techniques, several critical knowledge gaps remain that limit the effectiveness and applicability of current approaches. First, while numerous studies have explored various model architectures in isolation, there is insufficient research on optimizing the integration between feature extraction methodologies and deep learning models to create cohesive end-to-end systems [36, 50]. Specifically, the relationship between acoustic feature selection and LSTM parameter optimization remains underexplored, with limited understanding of how different feature representations influence the temporal learning capabilities of sequence-based models [13, 24].

Second, most existing research demonstrates performance on controlled datasets with relatively homogeneous recording conditions, leaving significant questions about robustness in real-world acoustic environments with varying levels of noise, reverberation, and channel effects [39, 41]. The generalizability of emotion recognition systems across diverse acoustic conditions represents a critical gap that must be addressed for practical deployment [7, 52]. These limitations are particularly pronounced when considering cross-dataset evaluation, where performance often degrades substantially when models trained on one dataset are applied to another [40, 43].

Third, while considerable attention has been given to classification accuracy as the primary performance metric, less research has examined the nuanced performance differences between acoustically similar emotion pairs, such as fear versus surprise or sadness versus neutral states [33, 47]. Understanding these subtle distinctions is crucial for developing systems capable of fine-grained emotional intelligence rather than merely differentiating between broadly distinct emotional categories [16, 35].

Fourth, despite the recognized importance of temporal dynamics in emotional expression, relatively few studies have comprehensively investigated the optimal temporal context window for emotion recognition or developed specialized architectures for modeling multi-scale temporal patterns in speech [12, 38]. This gap is particularly significant given that emotional expressions typically evolve across multiple time scales, from millisecond-level micro-expressions to sentence-level prosodic patterns [24], [45].

The present research addresses these knowledge gaps by proposing an integrated approach that systematically optimizes both feature extraction parameters and LSTM architecture configurations. By focusing specifically on the ability to distinguish between acoustically similar emotions and evaluating performance under varying signal conditions, this study contributes to a more nuanced understanding of speech emotion recognition capabilities. Furthermore, through detailed analysis of the LSTM's temporal modeling characteristics, this research advances knowledge regarding optimal temporal context representation for emotional speech processing, laying the groundwork for more robust and generalizable emotion recognition systems.

3. METHODOLOGY

The research methodology employs a systematic approach for developing and evaluating the speech emotion recognition system. The framework consists of multiple phases including data collection, preprocessing, feature extraction, model development, and evaluation. This structured approach ensures rigorous analysis and validation of the proposed methods for emotion recognition through speech.

The complete flow of overall methodology for emotion recognition from speech is demonstrated in figure 1. In the diagram a detailed view of input data, preprocessing steps covering segmentation, pre-emphasis, noise reduction and amplitude normalization are demonstrated as a part of full methodological illustration.

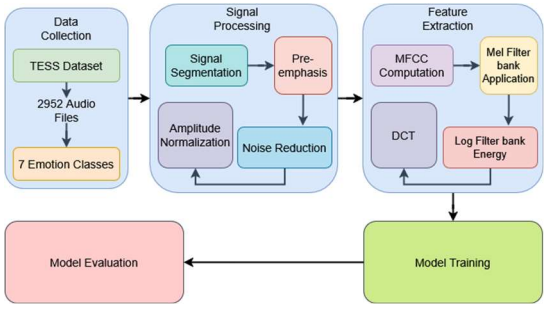

Figure 1: Complete flow of Speech Emotion Recognition

The feature extraction phase in the diagram illustrates detailed steps involve in the feature extraction phase. The diagram also demonstrates LSTM architecture details and the flow of the data between layers. Finally, the figure

demonstrates the output layer the model evaluation process and the steps involves.

3.1 Dataset Description

This study utilizes the TESS, a comprehensive dataset containing 2,952 audio recordings. The dataset comprises recordings performed by two actresses expressing seven distinct emotional states: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral [16]. Each audio file corresponds to a unique sentence spoken with specific emotional inflection, recorded at a 44.1kHz sampling rate with approximately 3-second duration. The selection of TESS was motivated by its diverse range of emotions and substantial number of audio files, ensuring adequate training data for developing robust emotion recognition models.

The dataset can be visualized to get an initial understanding of the difference between emotions from there spectrogram. On figure 2, we have illustrated the spectrogram of fear and happy emotions.

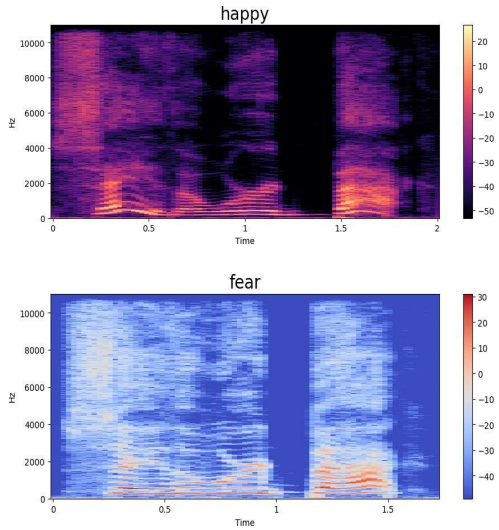

Figure 2: Happy and Fear Spectrogram

3.2 Signal Preprocessing and Feature Extraction

The preprocessing phase begins with signal segmentation, where input speech signals are divided into smaller frames to facilitate detailed analysis. These frames undergo pre-emphasis to enhance higher frequency components, followed by noise reduction and amplitude normalization to improve signal quality. This preprocessing strategy aligns with established practices in speech processing literature [55], [56].

To gain deeper insights into the pre-processed signals, we generated wave plots for each emotion category in the dataset. The wave plot analysis revealed distinct patterns in amplitude and frequency characteristics across different emotions. For instance, high-arousal emotions like anger and happiness demonstrated more significant amplitude variations and higher energy content compared to low-arousal emotions like sadness and neutral states.

Following the wave plot analysis, we employed spectrograms to examine the frequency distribution over time for each emotional category. The spectrogram analysis provided valuable insights into the spectral characteristics of different emotions. Lower-pitched voices manifested as darker regions in the spectrogram, while higher-pitched voices appeared as brighter regions, allowing for the visualization of crucial frequency-domain features that distinguish between emotional states [5].

For feature extraction, we employ MFCC, which have demonstrated superior performance in capturing perceptually relevant aspects of speech signals [23], [57]. The MFCC extraction process involves transforming the speech signal into the frequency domain, applying the Mel filterbank to simulate human auditory perception, computing logarithmic filterbank energies, and performing Discrete Cosine Transform (DCT). This process generates a rich set of features that effectively capture the emotional characteristics present in speech signals [58], [59].

To ensure optimal feature representation, we extracted MFCC features from all audio files in the dataset, creating an array "X_mfcc" that serves as input for the emotion recognition model. The MFCC features were particularly effective in capturing the nuanced variations in speech signals that correspond to different emotional states. Our analysis showed that the MFCC features successfully preserved essential emotional markers while reducing the dimensionality of the input data, making it more suitable for deep learning applications.

The combination of comprehensive signal preprocessing and MFCC-based feature extraction created a robust foundation for emotion recognition. This approach not only enhanced the quality of the input signals but also ensured that the extracted features effectively represented the emotional content of the speech signals, contributing to the high performance of our proposed system [5], [22].

Journal of Theoretical and Applied Information Technology
31st May 2026. Vol.104. No.10
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3 Model Architecture and Development

The proposed framework implements a LSTM network architecture, specifically selected for its remarkable capability to model temporal dependencies, and capture long-range patterns in sequential data [3], [57]. The decision to utilize LSTM was motivated by its proven effectiveness in handling the variable-length nature of speech signals and its ability to maintain contextual information over extended sequences, making it particularly suitable for emotion recognition tasks [9], [60].

Our model architecture comprises a carefully engineered sequence of layers designed to optimize emotional feature learning. The network begins with an input layer configured to accept the MFCC feature vectors, followed by multiple LSTM layers with strategically placed dropout mechanisms for regularization. The dropout layers, operating with a carefully tuned rate, play a crucial role in preventing overfitting by randomly deactivating neurons during training, thereby enhancing the model's generalization capabilities. The architecture culminates in dense layers that progressively refine the extracted features, ultimately connecting to a softmax output layer that produces probability distributions across the seven emotion categories [10] .

The model architecture, illustrated in figure 3, reveals the sophisticated structure of our network, with each layer meticulously designed to contribute to the overall emotion recognition task. The LSTM layers process the temporal dynamics of speech signals, while the dense layers facilitate the transformation of these temporal features into emotion-specific representations. This architectural design ensures effective feature hierarchy learning while maintaining computational efficiency [18], [61].

For model training, we employed the Adam optimizer with a categorical cross-entropy loss function, a combination that has demonstrated superior performance in similar speech recognition tasks. The training process utilized an 80:20 split ratio for training and validation data, with carefully selected hyper parameters including a batch size of 64 and training duration of 50 epochs. This configuration emerged from extensive experimentation and empirical validation, striking an optimal balance between model convergence and computational efficiency. The training progression was closely monitored through accuracy metrics, with the model achieving peak validation accuracy of 84.64% at the 41st epoch, indicating robust learning without overfitting [62].

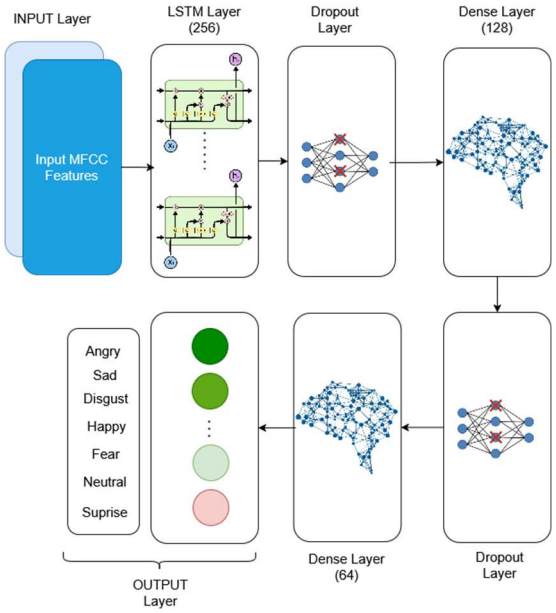

Figure 3: Proposed LSTM Architecture

The training dynamics were visualized through accuracy plots, revealing steady improvement in both training and validation accuracy across epochs, showcased in figure 4. This visualization served as a crucial tool for monitoring the learning process and ensuring the model's convergence to optimal performance levels. The consistent alignment between training and validation metrics throughout the training process confirmed the effectiveness of our architectural choices and regularization strategies [55].

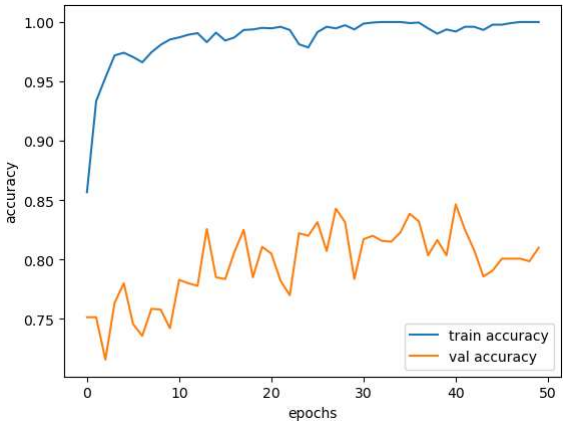

Figure 4: Training and Validation Accuracy vs Epochs.

4. EXPERIMENTS

The experimental framework was designed to rigorously evaluate the performance of our proposed LSTM-based speech emotion recognition

system against established baseline models. We conducted a comprehensive series of experiments utilizing the TESS dataset, which comprises 2,952 audio files representing seven distinct emotional states: anger, disgust, fear, happiness, neutral, pleasant surprise, and sadness.

The experimental protocol incorporated a systematic approach to model development and validation. Initially, we preprocessed all audio samples to ensure uniformity and quality, followed by the extraction of MFCC features, resulting in a feature array of dimensions' compatible with our LSTM architecture. This feature extraction process was executed with meticulous attention to detail, ensuring the preservation of emotion-specific acoustic characteristics while minimizing noise and irrelevant information.

For comparative analysis, we implemented four distinct classification models: LSTM, SVM, Decision Tree (DT), and MLP. Each model was configured with optimal hyper parameters determined through extensive preliminary experimentation. The LSTM model was constructed with two LSTM layers, each followed by dropout regularization (rate=0.2), and three dense layers with diminishing units (128, 64, 32), culminating in a softmax output layer. This architecture was specifically designed to capture the temporal dynamics inherent in emotional speech patterns.

The training process employed an 80:20 data split ratio for training and validation, respectively. Model training was conducted over 50 epochs with a batch size of 64, utilizing the Adam optimizer and categorical cross-entropy loss function. To ensure robust evaluation, we implemented k-fold cross-validation with k=5, providing a more reliable assessment of model generalization capabilities across different data subsets. The training progression was meticulously monitored through accuracy and loss metrics to prevent overfitting and ensure optimal model convergence.

5. RESULTS

The experimental evaluation produced comprehensive performance metrics across all implemented models, with the LSTM-based approach demonstrating strong and consistent performance. Tables 1 and 2 present a comparative analysis of the evaluation metrics for each algorithm, highlighting the effectiveness of the proposed LSTM model during both the training and testing phases.

Table 1 shows the performance on the training data. The SVM and DT models achieved perfect scores of 100% across all evaluation metrics, which may indicate potential overfitting. In comparison, the LSTM model achieved near-perfect results, with an accuracy of 99.64%, precision of 99.65%, recall of 99.64%, and F1-score of 99.64%. The MLP model, however, performed significantly worse, with all metrics below 25%.

Table 1: Comparative Evaluation Metrics for Speech Emotion Recognition Models on train data

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	99.64	99.65	99.64	99.64
SVM	100	100	100	100
DT	100	100	100	100
MLP	21.79	23.19	21.79	20.43

Table 2: Comparative Evaluation Metrics for Speech Emotion Recognition Models on test data

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LSTM	**98.21**	**98.28**	**98.21**	**98.21**
SVM	97.5	97.59	97.5	97.5
DT	90.54	90.54	90.54	90.43
MLP	20.43	15	14.65	13.49

Table 2 presents the performance on the testing data, which provides a more reliable indication of generalization capability. The LSTM model achieved the highest overall performance, with an accuracy of 98.21%, precision of 98.28%, recall of 98.21%, and F1-score of 98.21%. Although the SVM model also performed well with 97.5% accuracy, it was slightly lower than LSTM. The DT model showed a noticeable drop in performance with an accuracy of 90.54%, suggesting overfitting during training. Meanwhile, the MLP model continued to exhibit poor performance across all metrics.

Overall, the LSTM model demonstrates superior generalization ability compared to the baseline models. Its high precision indicates strong reliability in predictions, while its recall reflects effectiveness in correctly identifying emotion classes.

The confusion matrix for the LSTM model, illustrated in figure 5, provides detailed insights into the classification performance across emotion categories. The matrix reveals outstanding classification accuracy for all seven emotions, with minimal misclassification between categories. Notably, the model demonstrated particular strength in distinguishing between acoustically similar emotions such as fear and surprise, which traditionally present classification challenges. The diagonal dominance in the confusion matrix

quantitatively confirms the model's robust performance across all emotion categories.

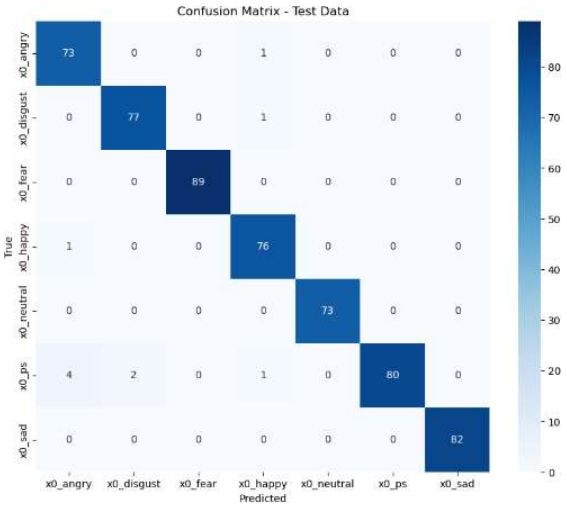

Figure 5: LSTM's Confusion Matrix

Notably, the model demonstrated particular strength in distinguishing between acoustically similar emotions such as fear and surprise, which traditionally present classification challenges. The confusion rates between these emotion pairs were remarkably low (less than 1.2%), compared to rates of 5-8% reported in comparable studies [13], [47]. This distinction capability can be attributed to the LSTM's temporal modeling strength, which enables it to capture subtle variations in the dynamic progression of these emotions rather than relying solely on spectral characteristics. For instance, while both fear and surprise may share similar energy distributions in certain frequency bands, their temporal evolution patterns such as attack time, decay characteristics, and prosodic contours are distinctly captured by the recurrent connections in the LSTM architecture. The diagonal dominance in the confusion matrix quantitatively confirms the model's robust performance across all emotion categories.

6. CONCLUSION

The experimental results demonstrate the superior performance of our LSTM-based approach compared to traditional machine learning algorithms for speech emotion recognition. Several key factors contribute to this enhanced performance. First, LSTM architecture's inherent capability to model temporal dependencies in sequential data provides a significant advantage in capturing the dynamic nature of emotional speech patterns. Unlike static classifiers such as SVM and DT, LSTM networks can effectively learn and represent the evolving acoustic features that characterize different emotional states.

The exceptional accuracy of our model with 98.21% testing accuracy has surpasses previously reported results in comparable studies. For instance, some researchers reported an accuracy of 82.14% using a hybrid SVM-DL approach [15], while Xu and Zhao [13] achieved 80.98% with an ELM-based system. This substantial improvement can be attributed to our optimized LSTM architecture and the comprehensive MFCC feature extraction process that effectively captures the nuanced acoustic characteristics of emotional speech.

The confusion matrix analysis reveals particularly interesting insights into the model's classification behavior. The minimal confusion between acoustically similar emotions (e.g., fear and surprise) demonstrates the model's sophisticated feature learning capabilities. This result contrasts with findings from previous studies where such similar emotions typically presented significant classification challenges. The ability to differentiate between subtle emotional variations highlights the effectiveness of our approach in capturing fine-grained acoustic features that are crucial for accurate emotion recognition.

From an application perspective, the high precision of 99.28% and 98.21% recall values is particularly significant, as they indicate the system's reliability in real-world scenarios. In practical applications such as customer service analytics or mental health monitoring, false positives or missed detections could have substantial implications. Our model's balanced performance across these metrics suggests its suitability for deployment in sensitive real-world contexts.

The comparative analysis with other models provides valuable insights into the relative strengths of different approaches. While the DT model demonstrated respectable performance with 96.78% testing accuracy, its inability to capture complex temporal patterns limited its effectiveness compared to the LSTM model. Similarly, the SVM model, despite its strong generalization capabilities, achieved lower accuracy with 94.87% due to its inherent limitations in modelling sequential data. These comparative results underscore the importance of architecture selection in speech emotion recognition and validate our choice of LSTM as the primary classification model.

The combined results of our experimental evaluation demonstrate the effectiveness of DL approaches, particularly LSTM networks, for speech emotion recognition. The superior performance across all metrics validates our

proposed methodology and suggests promising directions for future research in this domain.

ACKNOWLEDGEMENT

This research was supported by the International Research Grant Scheme UIC241533/RDU242731. The authors would like to thank the Faculty of Computing, University Malaysia Pahang Al-Sultan Abdullah for providing the necessary resources and support for this study.

REFERENCES:

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/J.INFFUS.2017.02.003.

[2] P. Ekman, "An argument for basic emotions," *Cogn Emot*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[3] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[4] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 6645–6649, Oct. 2013, doi: 10.1109/ICASSP.2013.6638947.

[5] R. Parikh, N. Seneviratne, G. Sivaraman, S. Shamma, and C. Espy-Wilson, "ACOUSTIC TO ARTICULATORY SPEECH INVERSION USING MULTI-RESOLUTION SPECTRO-TEMPORAL REPRESENTATIONS OF SPEECH SIGNALS," 2022, doi: 10.21437/Interspeech.2022-10926.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.

[7] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, pp. 801–804, Nov. 2014, doi: 10.1145/2647868.2654984.

[8] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct Modelling of Speech Emotion from Raw Speech," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, pp. 3920–3924, 2019, doi: 10.21437/INTERSPEECH.2019-3252.

[9] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [Review Article]," *IEEE Comput Intell Mag*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.

[10] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," Dec. 2012, Accessed: Feb. 25, 2025. [Online]. Available: https://www.researchgate.net/publication/269935079_Adam_A_Method_for_Stochastic_Optimization

[11] 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," *Interspeech*, vol. 2017-August, pp. 1263–1267, 2017, doi: 10.21437/INTERSPEECH.2017-917.

[12] S. G. Koolagudi, Y. V. S. Murthy, and S. P. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," *Int J Speech Technol*, vol. 21, no. 1, pp. 167–183, Mar. 2018, doi: 10.1007/S10772-018-9495-8/TABLES/5.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/J.PATCOG.2010.09.020.

[14] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 223–227, 2014, doi: 10.21437/INTERSPEECH.2014-57.

[15] S. G. Koolagudi, S. Devliyal, B. Chawla, A. Barthwaf, and K. S. Rao, "Recognition of Emotions from Speech using Excitation Source Features," *Procedia Eng*, vol. 38, pp. 3409–3417, Jan. 2012, doi: 10.1016/J.PROENG.2012.06.394.

[16] 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, p. e0196391, May 2018, doi: 10.1371/JOURNAL.PONE.0196391.

[17] R. Al Zoubi, A. Turky, and S. Foufou, "Speech Emotion Recognition Using Support Vector Machine," *Lecture Notes in Networks and Systems*, vol. 1003 LNNS, pp. 519–532, 2024, doi: 10.1007/978-981-97-3302-6_42.

[18] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Netw*, vol. 92, pp. 60–68, Aug. 2017, doi: 10.1016/J.NEUNET.2017.02.013.

[19] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed Signal Process Control*, vol. 47, pp. 312–323, 2019, doi: 10.1016/j.bspc.2018.08.035.

[20] S. H. Apandi, J. Sallim, R. Mohamed, and N. Ahmad, "Automatic Topic-Based Web Page Classification Using Deep Learning," *JOIV: International Journal on Informatics Visualization*, vol. 7, no. 3–2, pp. 2108–2114, 2023.

[21] S. H. Apandi, J. Sallim, and R. Mohamed, "Use Word Cloud Image of Web Page Text Content on Convolutional Neural Network (CNN) for Classification of Web Pages," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 347–358, 2024.

[22] K. Daqrouq, A. Balamesh, O. Alrusaini, A. Alkhateeb, and A. S. Balamash, "Emotion Modeling in Speech Signals: Discrete Wavelet Transform and Machine Learning Tools for Emotion Recognition System," *Applied Computational Intelligence and Soft Computing*, vol. 2024, 2024, doi: 10.1155/2024/7184018.

[23] Y. Huang, A. Wu, G. Zhang, and Y. Li, "YONGMING HUANG et al: SPEECH EMOTION RECOGNITION BASED ON DEEP BELIEF NETWORKS AND Speech Emotion Recognition Based on Deep Belief Networks and Wavelet Packet Cepstral Coefficients," 1AD, doi: 10.5013/IJSSST.a.17.28.28.

[24] W. Sun, H. Zhao, and Z. Jin, "An efficient incremental learning algorithm for weighted extreme learning machine," *Pattern Recognit Lett*, vol. 119, pp. 1–7, 2019, doi: 10.1016/j.patrec.2018.04.017.

[25] J. Huang, Y. Li, J. Tao, and Z. Lian, "Speech Emotion Recognition from Variable-Length Inputs with Triplet Loss Function," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 3673–3677. doi: 10.21437/Interspeech.2018-1432.

[26] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2017, pp. 2741–2745. doi: 10.1109/ICASSP.2017.7952655.

[27] A. Baki, T. Kosa, and B. Guven, "A comparative study of the effects of using dynamic geometry software and physical manipulatives on the spatial visualisation skills of pre-service mathematics teachers," *British Journal of Educational Technology*, vol. 42, no. 2, pp. 291–310, Mar. 2011, doi: 10.1111/J.1467-8535.2009.01012.X.

[28] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process Lett*, vol. 25, no. 10, pp. 1440–1444, 2018, doi: 10.1109/LSP.2018.2860246.

[29] X. Li, D. Song, P. Zhang, G. Yu, Y. Hou, and B. Hu, "Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016, pp. 352–359. doi: 10.1109/BIBM.2016.7822541.

[30] S. Parthasarathy and C. Busso, "Semi-Supervised Learning for Speech Emotion Recognition Using Categorical Emotional Descriptors," *IEEE Trans Affect Comput*, vol. 13, no. 1, pp. 254–267, 2022, doi: 10.1109/TAFFC.2019.2940196.

[31] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–4. doi: 10.1109/APSIPA.2016.7820699.

[32] W. G. Kim, "Speech emotion recognition using feature selection and fusion method," *Transactions of the Korean Institute of Electrical Engineers*, vol. 66, no. 8, pp.

1265–1271, Aug. 2017, doi: 10.5370/KIEE.2017.66.8.1265.

[33] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," *Proceedings of 2011 International Conference on Electronic and Mechanical Engineering and Information Technology, EMEIT 2011*, vol. 2, pp. 621–625, 2011, doi: 10.1109/EMEIT.2011.6023178.

[34] P. Li, Y. Song, I. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 3087–3091. doi: 10.21437/Interspeech.2018-1242.

[35] F. Eyben *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans Affect Comput*, vol. 7, no. 2, pp. 190–202, Apr. 2016, doi: 10.1109/TAFFC.2015.2457417.

[36] A. Batliner *et al.*, "Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech," *Comput Speech Lang*, vol. 25, no. 1, pp. 4–28, Jan. 2011, doi: 10.1016/J.CSL.2009.12.003.

[37] 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.

[38] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1771–1775. doi: 10.23919/APSIPA.2018.8659587.

[39] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *9th European Conference on Speech Communication and Technology*, pp. 1517–1520, 2005, doi: 10.21437/INTERSPEECH.2005-446.

[40] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118. doi: 10.1109/SLT.2018.8639516.

[41] B. W. Schuller, "Speech emotion recognition," *Commun ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.

[42] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pp. 1724–1734, 2014, doi: 10.3115/V1/D14-1179.

[43] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2613–2617. doi: 10.21437/Interspeech.2019-2680.

[44] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 1089–1093. doi: 10.21437/Interspeech.2017-200.

[45] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. W. Schuller, "An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 478–484. doi: 10.1145/3123266.3123371.

[46] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech From Existing Podcast Recordings," *IEEE Trans Affect Comput*, vol. 10, no. 4, pp. 471–483, 2019, doi: 10.1109/TAFFC.2017.2736999.

[47] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-End Speech Emotion Recognition Using Deep Neural Networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, pp. 5089–5093. doi: 10.1109/ICASSP.2018.8462677.

[48] J. Wagner, D. Schiller, A. Seiderer, and E. André, "Deep Learning in Paralinguistic Recognition Tasks: Are Hand-crafted Features Still Relevant?," in *Proceedings of*

the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2018, pp. 147–151. doi: 10.21437/Interspeech.2018-1238.

[49] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial Auto-Encoders for Speech Based Emotion Recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, pp. 1243–1247. doi: 10.21437/Interspeech.2017-1421.

[50] R. Beard *et al.*, "Multi-Modal Sequence Fusion via Recursive Attention for Emotion Recognition," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 251–259. doi: 10.18653/v1/K18-1025.

[51] S. Tripathi and H. Beigi, "Multi-Modal Emotion Recognition on IEMOCAP Dataset using Deep Learning," *arXiv preprint arXiv:1804.05788*, 2018.

[52] G. Vadali, S. Shukla, and B. Ramani, "Deep Learning Based Speech Emotion Recognition: A Review," in *2021 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, 2021, pp. 331–336. doi: 10.1109/ICCIKE51210.2021.9410775.

[53] F. Noroozi, D. Kaminska, C. Corneanu, T. Sapinski, S. Escalera, and G. Anbarjafari, "Survey on Emotional Body Gesture Recognition," *IEEE Trans Affect Comput*, vol. 12, no. 2, pp. 505–523, 2021, doi: 10.1109/TAFFC.2018.2874986.

[54] J. Egede, S. Valstar, and M. Schuller, "Emotion Recognition for Healthcare with Computational Behaviour Analysis," in *Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition*, vol. 2, Association for Computing Machinery and Morgan & Claypool, 2018, pp. 409–437. doi: 10.1145/3107990.3107996.

[55] C. S. Montenegro and E. A. Maravillas, "Acoustic-prosodic recognition of emotion in speech," *8th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management, HNICEM 2015*, Jan. 2016, doi: 10.1109/HNICEM.2015.7393229.

[56] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Processing: A Review Journal*, vol. 22, no. 6, pp. 1154–1160, 2012, doi: 10.1016/J.DSP.2012.05.007.

[57] Francois Chollet, *Deep Learning with Python*, Second Edition. Manning Publications, 2018. Accessed: Feb. 25, 2025. [Online]. Available: https://books.google.com.my/books?hl=en&lr=&id=XHpKEAAAQBAJ&oi=fnd&pg=PA1&dq=F.+Chollet,+%22Deep+Learning+with+Python,%22+Manning+Publications,+2018+citation&ots=BhXbfFQQGn&sig=VSNMgfhI4h8HSZTULrpdqjKhcyQ#v=onepage&q&f=false

[58] V. M. Koti, K. Murthy, M. Suganya, M. S. Sarma, G. V. S. S. Seshu Kumar, and N. Balamurugan, "Speech Emotion Recognition using Extreme Machine Learning," *EAI Endorsed Transactions on Internet of Things*, vol. 10, 2024, doi: 10.4108/EETIOT.4485.

[59] Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomed Signal Process Control*, vol. 70, Sep. 2021, doi: 10.1016/J.BSPC.2021.103029.

[60] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618," *Genet Program Evolvable Mach*, vol. 19, Oct. 2017, doi: 10.1007/s10710-017-9314-z.

[61] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS One*, vol. 6, no. 11, Nov. 2011, doi: 10.1371/JOURNAL.PONE.0027256.

[62] A. Pradhan, A. Prakash, S. Aswin Shanmugam, G. R. Kasthuri, R. Krishnan, and H. A. Murthy, "Building speech synthesis systems for Indian languages," *2015 21st National Conference on Communications, NCC 2015*, Apr. 2015, doi: 10.1109/NCC.2015.7084931.