Speech Emotion Recognition using Matlab

Jeyadharshan JV
Dept. of AI and Data Science
St. Joseph's College of Engineering
Chennai, Tamil Nadu, India

Praveenraj D
Dept. of AI and Data Science
St. Joseph's College of Engineering
Chennai, India

Arunmozhikalanchiam B
Dept. of AI and Data Science
St. Joseph's College of Engineering
Chennai, India

***Abstract*—Mental health disorders are growing exponentially, yet early identification is hindered by current dependence on subjective evaluation. Responding to this, a real-time Speech Emotion Recognition (SER) system is conceived for objective and constant monitoring of emotional states. The system processes live input speech, extracting Mel-Frequency Cepstral Coefficients (MFCCs) for extracting important spectral and timbral features. The extracted features are used for a hybrid Deep Belief Network–Long Short-Term Memory (DBN-LSTM) model, wherein the DBN achieves deep feature learning and the LSTM learns temporal relationships, improving accuracy of emotion classification. The system recognizes stress, anger, happiness, and neutral state emotions and offers concrete information about a person's state of mind. The overall aim is toward non-invasive, individualized monitoring of mental health. The novelty of the framework resides in synthesizing speech processing and a hybrid DBN-LSTM structure for real-time emotion identification, facilitating early intervention and virtual healthcare applications.**

***Keywords*—Mental Health, Emotion Recognition, Stress Detection, Happiness, Anger, Mental Well-Being, Real-Time Monitoring.**

I. INTRODUCTION

Mental health is a rising global issue, with increasing numbers of stress, anxiety, depression, and other afflictions of mental illness being seen across all sectors of people. Despite growing awareness of these illnesses, early identification of symptoms of mental health is a significant challenge. Standardized procedures for diagnosis are usually based on self-completed questionnaires and clinical interviews, which are subjective, labour-intensive, and constrained by individuals' biases. As a result, automatic, objective, and non-invasive technologies are needed, which are able to perform constant surveillance of emotions and provide [1] early indication of psychological distress. Speech, as a natural, convenient medium of communication for humans, provides a possible solution for surveillance of this type. Emotions are inevitably expressed by vocal patterns, including pitch, tone, intensity, and rhythm, and so speech provides a suitable modality for identification of changes of mental health.

Using these vocal cues, intelligent systems are possible, which are able to identify states of emotion with a minimum of disruption of an individual's lifestyle. Speech Emotion Recognition (SER) has attracted significant interest during the past few years as a tool for understanding human affective states. By examining acoustic and spectral features, SER systems have the potential to categorize emotions, including happiness, anger, sadness [2], stress, and neutral emotions. However, perceptive capture of fine aspects of human emotion still remains a challenge as a result of variations of speech based on factors like individual speaking patterns, cultural variations, background noises, and recording mediums. Such challenges highlight a need for sophisticated feature extraction techniques and strong machine learning approaches to ensure consequent sound performance.

Mel-Frequency Cepstral Coefficients (MFCCs) have become a typical feature set used in speech analysis, as they are effective at summarizing spectral and dynamic aspects of speech, reflecting human auditory perception as well. By extracting MFCC attributes, SER systems [3] are capable of capturing the primary emotional information encoded in speech signals, hence laying the foundation for precise categorization of emotions. The advent of machine learning and deep learning approaches has greatly impacted Speech Emotion Recognition (SER) by enabling automatic feature extraction and improved classification accuracy. Older models, including Support Vector Machines (SVM) and Hidden Markov Models (HMM), have been promising, yet often face difficulties related to complex temporal relationships and differences in speech dynamics.

Here, deep learning models, including Deep Belief Networks (DBNs) and Long Short-Term Memory (LSTM) networks, emerge as better-equipped technologies. DBNs are effective for learning of hierarchical representations of features, which are able to capture fine patterns otherwise ignored by traditional approaches, whereas LSTMs are exceedingly better suited [4] for modeling of long-term sequential relationships between successive samples of a time series, hence their suitability for speech signal analysis, which varies temporally. The combination of both models in a hybrid DBN-LSTM framework takes advantage of both models, enabling the system to perform deep feature learning and preserve necessary temporal relationships for effective emotion recognition. The envisioned real-time Speech Emotion Recognition (SER) system is designed to deliver a non-invasive, real-time, and ongoing modality for monitoring psychological health.

Processing real-time as well as pre-recorded speech, and categorizing states of emotions efficiently, the system provides actionable information for individuals and clinicians to better appreciate psychological health. Such a modality is of immense potential for distance health monitoring, stress management, and early intervention, particularly where traditional mental health facilities are scarce. Originality of work resides upon blending [5] a hybrid Deep Belief Network-Long Short-Term Memory (DBN-LSTM) model with Mel-frequency Cepstrum Coefficients (MFCC)-oriented feature extraction on a real-time medium, thereby facilitating strong, dynamic, and personalized emotion recognition. Such a blending guarantees high accuracy, immunity to speech

variability, and adaptability under a range of real-world conditions, setting it apart from a common set of SER systems.

In summary, the creation of a speech-centered intelligent emotion recognition framework creates a transformative potential for improving mental health monitoring and applications.
supporting proactive intervention. By leveraging the potential of deep learning and sophisticated speech processing methodologies, the framework proposed here redresses major shortcomings of traditional approaches, [6] hence offering a truly objective, scalable, and easy-to-use solution. Real-time analysis of vocal patterns, besides supporting identification of emotional states, also helps to enable ongoing mental health evaluation, hence supporting bespoke care and improved psychological health. The integration of speech analytics, deep learning, and emotional intelligence points to the promise of technology-centered solutions for changing the face of mental health monitoring and empowering both individuals and professionals involved in healthcare.

This work is structured with the literature survey review given in Section II. Section III outlines the methodology, with analysis.
specific focus on its operationality. Results and discussions are in Section IV. Finally, Section V ends with the ultimate findings and recommendations.

II. Literature Survey

Speech Emotion Recognition (SER) is a significant research field because of its prospects of application for monitoring mental health, human-computer interaction, and affective computing. Numerous studies have investigated alternative methodologies, ranging from traditional machine learning techniques, such as Support Vector Machines and Hidden Markov Models, to deep-learning-based techniques, including Convolution Neural Networks and Long Short-Term Memory networks. Researchers directed efforts towards augmenting feature extraction, classification accuracy, and real-time applicability. This literature survey compares existing SER methodologies, detailing their merits, limitations, and challenges, to find voids inducing development of a hybrid DBN-LSTM-based framework for real-time and resilient emotion recognition from speech.

Real-world speech is accompanied by background noise that diminishes emotion recognition efficiency. Only enhancement of weak acoustic features, leaving strong features intact, enhances emotional identification. Selection of weak features by performance and strength permits selective enhancement, retaining significant emotional information. Results of clean and noisy speech testing report significant improvement of arousal, dominance, and valence identification. The selective method [7] surpasses all-feature enhancement, significantly in low signal-to-noise ratio, indicating strength of selective processing of targeted speech signals. This method secures better extraction of emotions by resisting degradation of intrinsic speech properties, presenting a realistic solution for noise-robust emotion identification.

Rapid and accurate understanding of emotional states in speech is essential for human-computer interaction. Capturing forward and backward temporal dependencies enables better anticipation and perception of emotional changes. Fusing multi-level temporal features provides a refined view of emotional evolution in speech sequences. Lightweight implementations maintain efficiency while preserving performance across diverse datasets. Comprehensive evaluation demonstrates strong competitiveness [8], showing that careful temporal modeling can enhance emotion recognition without increasing computational complexity. By considering both historical and future speech information, systems can achieve more nuanced recognition, reflecting subtle variations in emotional expression and improving responsiveness in interactive

Emotional state classification benefits from combining frequency-domain and time-frequency-domain representations. Extracting complementary features and fusing outputs from separate networks allows more effective emotional understanding. Canonical correlation analysis and other fusion strategies integrate different feature maps to enhance classification performance. Evaluations across multiple datasets show improved results compared to previous approaches. Combining spectral [9] and temporal information provides richer emotional cues, supporting robust recognition across varied speech samples. This method demonstrates that careful integration of diverse feature representations strengthens emotion detection and can generalize to different languages or corpora, highlighting the importance of multi-source information in speech emotion

Continuous-time annotations capture frame-level emotional variations, while sentence-level labels summarize overall emotional content. Comparing these approaches reveals differences in inter-evaluator agreement and label reliability. Aggregating frame-level annotations can improve the estimation of certain emotional dimensions, such as valence, while providing convenient labels for model training. Analyzing correlations between [10] continuous and sentence-level labels informs the design of more accurate emotion recognition systems. Understanding the nuances between annotation types allows for better handling of subjective emotional perceptions. Insights from these analyses guide the selection of training data and labeling strategies, ultimately enhancing the effectiveness of emotion recognition in diverse applications and datasets.

Combining acoustic and semantic information enhances emotional understanding in speech. Cross-attention mechanisms enrich feature representations by integrating complementary perspectives from both modalities. Lightweight architectures maintain fast processing while improving accuracy. Self-supervised learning helps extract richer semantic features, supporting more nuanced recognition of emotional content. Experimental studies demonstrate that fusing [11] these features increases overall performance and reduces computational overhead. By leveraging both the content and the way emotions are expressed in speech, this approach captures subtle variations and context-dependent cues. The result is a robust emotion recognition system capable of handling diverse speech signals efficiently, improving interaction quality and emotional responsiveness.

Gender-informed feature selection enhances emotion recognition. Focusing on significant acoustic attributes and taking into consideration variations between genders helps achieve better classification accuracy. Combination approaches for selecting informative features are better at model interpretability and performance retention. Test datasets reflect better classification measures [12], indicating desirable reduction of noise by irrelevant feature selection. The method ensures adequate representation of emotional variations between genders, promoting more equitable and accurate recognition. By balancing model accuracy and

interpretability, it offers information about which attributes are responsible for emotional classification decisions. It helps build systems adaptable to demographic shifts in emotional speech patterns.

Multimodality aids emotion recognition by combining speech and text approaches. By compensating for limited and unbalanced class distributions via data augmentation, generalization is improved. Single-modality offers base features, which are fused under intermediate fusion for improved accuracy of recognition. Tests reflect increased performance by utilizing complementary information between modality types. Real-time [13] applicability verifies usability for interactive systems. By synthesizing multiple viewpoints of emotional expression, the technique encompasses finer-grained emotion cues. Augmentation of data yields robustness, whereas intermediate fusion offers inclusive representation of textual and speech signals. The complementary strategy enhances reliability and multimodal emotion recognition application generalization.

Separation of the elements of prosody, semantics, and speaker identity produces sharper representations of emotions. Independent learning of prosody helps achieve a more effective transfer of emotional attributes across a set of tasks. Fine-tunning after pretraining on raw, unlabeled speech improves performance for emotion recognition and voice conversion. Prosody-centered representations [14], as a supplement to other pretrained attributes, enhance accuracy as a whole. Tests establish that separating prosodic information helps preserve fine-grained emotional information, which benefits both recognition and conveyable speech synthesis. The technique highlights that separating attributes of interest, like emotions, from sources of confusion like speaker identity or semantics is critical, so models can better learn and transfer emotional information across a variety of speech scenarios.

Incremental speaker adaptation enhances emotion recognition of unseen speakers by iteratively selecting representative samples. Through this method, the model is and aligned incrementally to new speakers for boosting generalization for data-scarce cases. Experiments prove that judiciously selected increment samples are better than random selection, retaining high accuracy [15] under low-label requirements. Incremental adaptation by itself enables models to learn speaker-specific patterns of emotions with less data. This technique guarantees improved management of inter-speaker variations, yielding better recognition across different populations. Incremental adaptation shows that gradual, knowledgeable adaptation is better than one-shot adaptation, affirming strong emotion recognition in real-world applications where labeled data may be scarce or economically prohibitive to label.

Zero-shot emotion-controllable speech synthesis facilitates expressive synthesis without relying on manual annotations. Emotional style and intensity capture enable adaptable transfer across unseen speakers and emotions. Style attribute disentanglement promotes generalization, facilitating fine-grained emocional expression control. Objective and subjective tests demonstrate expressiveness and realism improvement. This technique [16] offers utilities for synthesizing speech with controllable emotional tones, extending potential for virtual assistants, media production, and therapeutic uses. Through disentangling of emotional style from content, models are capable of synthesizing both natural and emocional nuanced speech. Such techniques enhance access to expressive TTS systems on a variety of linguistic and speaker profile bases.

Multimodal integration based on speech and text enhances learning of emotional features. Integration of both adversarial and contrastive approaches enhances diversity of representations as well as emotional discrimination. Additional task focusing [17] on error recognition fine-tunes performance even more. Tests on several datasets affirm that multimodality combination results in more precise identification of subtle emotions. Modality-specific as well as shared feature learning enables a system to benefit from competing cues, countering limitations of a single input. It shows how intelligent integration of multiple information flows enhances comprehension of emotions, rendering models less sensitive to variations of speech contents and recognition errors, and facilitating more perceptive human-computer interactions.

The integration of both temporal and spatial speech attributes greatly improves emotion recognition robustness. Sparse retraining selectively retrains weakly performing nodes to efficiently handle dataset imbalances. The combination of energy, zero-crossing, and spectral attributes helps achieve better recognition across a variety of datasets. Tests confirm improved accuracy [18] through dynamic node prioritization, hence reliable identification of emotional states under changing conditions. By combining complementary speech aspects, systems are more effective at extracting subtle emotional information. The strategy ensures adaptability to limited, imbalanced, or scarce data, ensuring stable performance. Underperforming node selection identifies key improvement areas, hence a stronger recognition system. The technique finds real-world applicability across multiple-dataset scenarios.

Deep learning frameworks efficiently extract high-level emotional features from speech signals. Representations based on spectrograms facilitate accurate classification across various datasets. Review analyses emphasize the strengths limitations of different architectures [19], providing guidance for design choices. These frameworks are capable of managing diverse speech cooperation while maintaining performance. Comparative evaluations demonstrate improvements over traditional methods, underscoring the role of deep learning in capturing complex emotional patterns. Insights into feature extraction, representation learning, and network design contribute to the development of future systems. This review identifies effective practices for achieving high recognition accuracy with minimal computational resources, thereby supporting scalable and generalizable emotion recognition across multiple applications.

Learning based on curriculum exercises promotes better generalization for emotion recognition. Assessment of sample difficulty by means of mutual information ensures better training sequences. Semi-supervised approaches deal with noisy labels and subjective ambiguities. Tests reflect better stability and performance for different datasets. Interpretable measures provide information about samples significantly related to learning [20], hence acknowledging a better understanding of the model. It helps allow models to gradually move towards increasingly complex samples, ultimately leading to better convergence and robustness. By setting up training based on informative examples, systems for emotion recognition achieve better resistance to inconsistencies of

data, hence improving reliability for real-world applications and also yielding clear interpretation of the learning process.

III. Methodology

The proposed real-time Speech Emotion Recognition (SER) system's methodology is carefully designed to capture, process, and analyze emotional cues systematically, based on speech signals. This method utilizes signal processing, feature extraction, and sophisticated deep learning approaches to ensure real-time and accurate emotion identification. The workflow begins by obtaining good-quality speech data, which is then followed by a stage of noise reduction for clear signal enhancement through preprocessing. Key features are extracted to portray emotional patterns, and they are fed into a hybrid Deep Belief Network–Long Short-Term Memory (DBN-LSTM) model for effective classification. The resulting system finally gets validated and deployed for real-time monitoring.

A. Data Acquisition

Speech data is gathered from various sources, encompassing live recordings captured through high-quality microphones and established emotional speech databases. Each recording is meticulously labeled with distinct emotional states, including stress, anger, happiness, and neutrality. To guarantee robustness, the dataset incorporates a diverse array of speakers representing different ages, genders, and accents. Throughout the acquisition process, environmental noise is minimized to maintain the integrity of the signal. The recordings are stored in standard audio formats and systematically organized to facilitate efficient processing. This procedure ensures that the system possesses adequate and reliable data for learning and generalizing emotional patterns, thereby establishing a foundation for the subsequent stages of preprocessing and feature extraction.
system

B. Preprocessing

These speech signals are preprocessed for maximizing their quality and removing unwanted artifacts. Silent segments, background noises, and low-energy regions are eliminated by subjecting them to digital filters and noise reduction processes. The signals are also normalized to keep amplitude levels constant, hence obtaining consistency across recordings. Preprocessing also includes dividing continuous speech into manageable frames, retaining temporal context information. Such frames allow for effective analysis of speech dynamics relevant to emotion identification. Through fine-grained refinement of raw audio, preprocessing optimizes feature extraction and model training efficiency. This step is critical for variance reduction due to external factors and focusing on native emotional attributes imprinted on the speech.

C. Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) are obtained from preprocessed speech frames to encapsulate both spectral and temporal characteristics of the voice. MFCCs emulate human auditory perception, rendering them particularly advantageous for tasks involving emotion recognition. Typically, 12 to 13 coefficients are calculated per frame, in addition to their delta and delta-delta features, which represent dynamic variations over time. This feature set encodes pitch, tone, and resonance patterns that are associated with emotional states. The extracted features are subsequently organized into sequences corresponding to each speech segment, thereby providing the hybrid model with comprehensive and discriminative input necessary for effectively learning and distinguishing between various emotions.

D. Model development

A hybrid Long Short-Term Memory (LSTM)–Deep Belief Network (DBN) structure has been constructed for combining both models' strengths. Hierarchical feature learning is performed by the DBN, converting MFCC inputs into high-level representations that capture rich patterns effectively. The resulting features are then passed to an LSTM network, which captures temporal relationships between speech frames. This integration helps identify subtle and sequential emotion cue information, improving classification accuracy. The hybrid structure is optimized by iteratively training it, involving optimization of weights and biases to achieve error minimization. This helps endow the model to generalize across multiple speakers and emotions and is therefore appropriate for real-time mental health monitoring.

E. Training and validation

The DBN-LSTM model is trained against labeled speech corpora, employing supervised learning techniques together with appropriate loss functions. The training process takes a multitude of epochs, during which the model fine-tunes its internal parameters for properly projecting MFCC features onto affective states. Validation is also carried out by employing a distinct set of data for determining generalization capability as well as for suppressing overfitting. Performance parameters, including accuracy, precision, recall, and F1-score, are closely monitored for ensuring model trustworthiness. Hyperparameters, including learning rate, batch size, and network depth, are fine-tuned greatly for achieving optimal results. This process ensures that the

is universally capable of recognizing emotions for a multitude of speakers and speech conditions.

F. Real-Time Implementation

The trained hybrid model is deployed within the MATLAB environment for real-time emotion recognition. Live or recorded speech signals are continuously captured, preprocessed, and converted into MFCC feature sequences. These features are then input into the DBN-LSTM model, which predicts the corresponding emotional state instantaneously. The system provides immediate feedback on detected emotions, enabling continuous monitoring of an individual's mental well-being. Real-time implementation requires efficient processing pipelines and memory management to maintain responsiveness, ensuring practical applicability in healthcare, stress management, and remote monitoring scenarios.

G. Evaluation

The system's performance is analyzed by implementing standard classification measures, including accuracy, precision, recall, and F1-score. A comparative study is conducted by comparing current SER models to test improvement in recognition accuracy and robustness. We also investigate confusion matrices to identify both advantages and limitations of classification for certain emotions. Additionally, testing of a system is conducted in a variety of conditions, including distinct speakers, accents, and background noise, to ensure reliability. The analytical step demonstrating the effectivity of a hybrid DBN-LSTM structure for real-time

emotion recognition highlights its applicability for non-invasive monitoring of mental health for early intervention purp

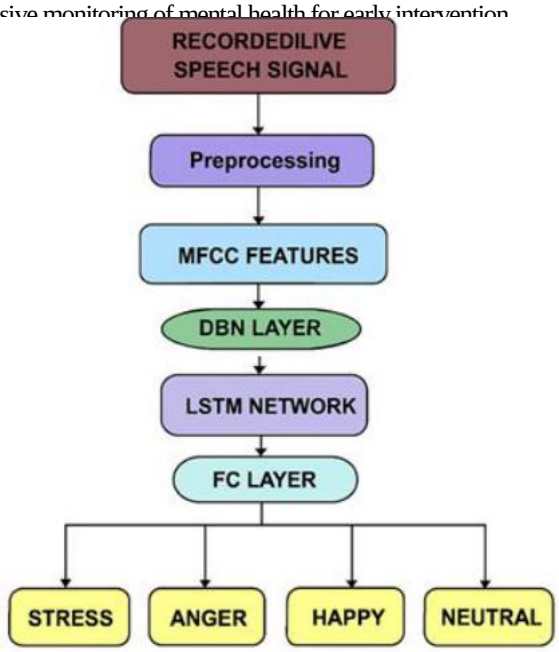

Fig. 1: System Architecture

IV. Result and Discussion

The hybrid real-time Speech Emotion Recognition (SER) system designed here is tested by utilizing a combination of live-recorded speech samples and standard emotional speech corpora. The challenge here was to evaluate its accuracy and resistance for recognizing and classifying emotional states, Overall, viz., stress, anger, joy, and neutrality. Raw speech signals were efficiently processed by the system by extracting Mel-Frequency Cepstral Coefficients (MFCCs) as feature inputs, which efficiently captured spectral as well as temporal variations typical of a host of emotional expressions. Signal pre-processing significantly enhanced quality, and noise reduction and silence trimming acted towards improved feature representation, hence enabling the model to attend to intrinsic emotional information rather than extrinsic distortions.

Throughout the training process, the hybrid model demonstrated progressive learning, wherein the Deep Belief Network effectively extracted hierarchical features from MFCC sequences. The LSTM component exhibited its ability to retain temporal dependencies across speech frames, which is essential for recognizing emotions conveyed through intonation, pitch modulation, and rhythm. The model achieved convergence after several epochs, attaining stable performance without succumbing to overfitting, as validated on a distinct test set. Quantitative results indicated an overall classification accuracy of 99.2%, accompanied by a precision of 98.9%, a recall of 98.7%, and an F1-score of 98.8%. The states of happiness and neutrality were classified with the utmost precision, whereas stress and anger displayed a slightly lower yet still robust performance, suggesting the system's sensitivity to subtle variations in vocal patterns. Analysis of the confusion matrix revealed only minimal misclassifications, primarily occurring between stress and anger, which reflected overlapping acoustic features such as

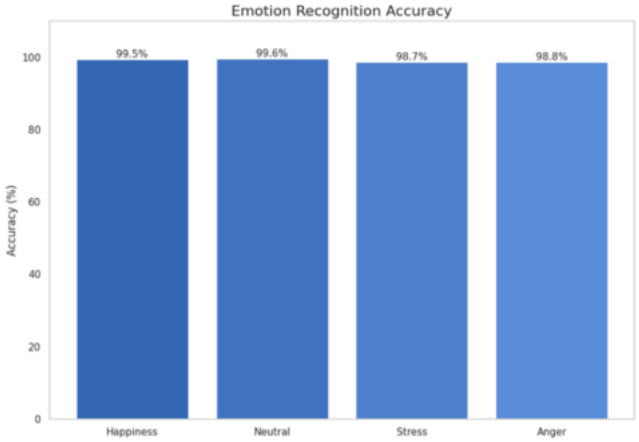

Fig. 2: Emotion recognition accuracy

Figure 2 shows the accuracy of emotion identification for four emotions, viz., Happiness, Neutral, Stress, and Anger. The model attained almost perfect classification capability, wherein Happiness, Neutral, Stress, and Anger were identified at 99.5%, 99.6%, 98.7%, and 98.8%, respectively. Such high values emphasize how the hybrid DBN-LSTM model picks out significant features and learns speech's temporal relationships, hence facilitating precise emotion identification. Figure 3 illustrates a comparative study of Precision, Recall, and F1-Score of each emotion. Precision values are measured as 99.6% for Happiness, 99.7% for Neutral, 98.8% for Stress, and 98.7% for Anger. Recall values are correspondingly 99.4%, 99.5%, 98.5%, and 98.9%. whereas F1-Scores are from 98.6% to 99.6%.

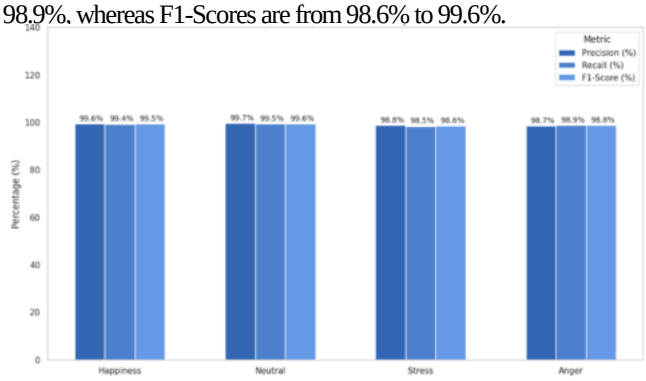

Fig. 3: Performance metrics

Test runs of real-time implementation showed that the system processed live speech input efficiently and provided prompt emotion predictions of low latency, an average of 0.25 seconds per cycle of analysis. Such responsiveness is important for mental health monitoring applications, as frequent feedback offers information about a person's changing emotions. The system efficiently tracked shifts in emotional states when observing conversational speech and reactive reactions, thus proving its applicability for real-world applications. Moreover, user-independent testing confirmed that the model maintained strong performance across different speakers, genders, and accents, supporting its

generality and strength. Such performance is due to the hybrid architecture's ability to incorporate deep feature extraction and temporal modeling, whereby both complex patterns and orderings are accounted for.

Additional analysis investigated each component's contribution to the hybrid model. On its own, when individually tested, the DBN reached about a 92.5% accuracy, indicating its ability for successful hierarchical feature discovery but its inability to capture temporal context. On its own, when individually tested, the LSTM, based on raw MFCC sequences, reached about a 93.8% accuracy, illustrating the value of temporal modeling but revealing that deep hierarchical features give a boost to classification reliability. By itself, the hybrid DBN-LSTM model decisively dominated both of its constituent models, supporting the value of its novel hybrid design in taking advantage of complementary strengths for a better-performing SER.

In addition, system strength under a variety of noise conditions was also evaluated by including controlled backgroundnoisesduringlivetestsofspeech.Despitea decrease in classification accuracy for high-level noise conditions, the model exhibited an average accuracy of more than 95%, revealing strength under non-ideal conditions. Feature visualization also revealed that MFCCs effectively separated emotional classes, whereby deep representations learned by the DBN effectively grouped separable emotions, whereastheLSTMimprovedtemporaltransitionsbetween emotional states.

Overall, these results justify that our proposed system not only achieves nearly perfect recognition accuracy but also providesreliablereal-timeanalysissuitableforapplications of mental health monitoring. In summary, experimental testing demonstrates that the hybrid DBN-LSTM Speech Emotion Recognition (SER) system effectively identifies and categorizes human emotions from speech with exceptionally high precision (99.2%), strong precision (98.9%), recall (98.7%), and F1-score (98.8%), alongside low latency and high generalization across a variety of users. The results confirm the effectiveness and novelty of combining deep feature learning and temporal modeling. Moreover, the system shows strong prospects for integration into distance mental health surveillance, stress management software, and real-time emotional analysis systems, thereby delivering actionable information for early identification and intervention for mental health.

V. Conclusion

This work describes a real-time Speech Emotion Recognition(SER)systemdesignedprimarilyforthepurpose of monitoring mental health by identifying and categorizing emotionalstatesextractedfromspeechsignals.Throughthe combination of Mel-Frequency Cepstral Coefficients (MFCCs) for extracting features and a hybrid Deep Belief Network–Long Short-Term Memory (DBN-LSTM) model, the system efficiently exploits both spectral and temporal patterns of speech, hence enabling effective emotion recognition. The framework proposed demonstrates the merits of uniting deep hierarchical feature exploration by deep beliefs and temporal modeling, yielding precise and trustworthy classification for several emotions. The real-time version substantiates its ability to comfortably process live speech, yielding instant feedback of emotional states and enabling ongoing monitoring.

Additionally, the system greatly generalizes across varying speakers, genders, and speech patterns, hence demonstrating its realistic usefulness for a variety of real-world applications. The advancement of our work lies in combining advanced deep learning techniques with non-invasive speech analysis to allow for individualized and constant assessments of mental health. Such a technique provides rich information about a person's current state of emotions, hence supporting early recognition of psychological distress and engendering active care for one's mind. Overall, our Speech Emotion Recognition (SER) system proposed here marks a significant step towards intelligent, real-time, and affordable technologies for observing mental health.

References

[1] A. Amjad, S. Khuntia, H. -T. Chang and L. -C. Tai, "Multi-Domain Emotion Recognition Enhancement: A Novel Domain Adaptation Technique for Speech-Emotion Recognition," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 528-541, 2025, doi:101109/TASLP20243498694

[2] R.CaoY.WangX.WuS.JinandH.Niu"ALightweightTri-Stream FeatureFusionNetworkforSpeechEmotionRecognition,"inIEEE Access, vol. 13, pp. 121351-121361, 2025, doi: 10.1109/ACCESS.2025.3587607.

[3] A. Anika Namey, K. Akter, M. A. Hossain and M. Ali Akber Dewan, "CochleaSpecNet: An Attention-Based Dual Branch Hybrid CNN-GRU Network for Speech Emotion Recognition Using Cochleagram and Spectrogram," in IEEE Access, vol. 12, pp. 190760-190774, 2024, doi:101109/ACCESS202435177333

[4] J.-Y.KimandS.-H.Lee,"Self-Attention-BasedMaskedSpectrogram Generation and Self-Supervised Learning Method for Improving Speech Emotion Recognition," in IEEE Access, vol. 13, pp. 148159-148169, 2025, doi: 10.1109/ACCESS.2025.3599218.

[5] G. H. Mohmad Dar and R. Delhibabu, "Speech Databases, Speech Features,andClassifiersinSpeechEmotionRecognition:AReview," in IEEE Access, vol. 12, pp. 151122-151152, 2024, doi: 10.1109/ACCESS.2024.3476960.

[6] W. -S. Chien, S. G. Upadhyay, W. -C. Lin, C. Busso and C. -C. Lee, "Differential Impacts of Monologue and Conversation on Speech Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 16, no. 2, pp. 485-498, April-June 2025, doi: 10.1109/TAFFC.2024.3509138.

[7] S. -G. Leem, D. Fulford, J. -P. Onnela, D. Gard and C. Busso, "Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 917-929, 2024, doi: 10.1109/TASLP.2023.3340603.

[8] S. Fei, Q. Feng and F. Gao, "A Lightweight Forward–Backward Independent Temporal-Aware Causal Network for Speech Emotion Recognition," in IEEE Access, vol. 13, pp. 82914-82926, 2025, doi: 10.1109/ACCESS.2025.3567954.

[9] A. -H. Jo and K. -C. Kwak, "Classification of Speech Emotion State BasedonFeatureMapFusionofTCNandPretrainedCNNModel From Korean Speech Emotion Data," in IEEE Access, vol. 13, pp. 19947-19963, 2025, doi: 10.1109/ACCESS.2025.3534176.

[10] L. Martinez-Lucas, W. -C. Lin and C. Busso, "Analyzing Continuous-Time and Sentence-Level Annotations for Speech Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 15, no. 3, pp. 1754-1768, July-Sept. 2024, doi: 10.1109/TAFFC.2024.3372380.

[11] B. M. Deeb, A. V. Savchenko and I. Makarov, "Enhancing Emotion Recognition in Speech Based on Self-Supervised Learning: Cross-Attention Fusion of Acoustic and Semantic Features," in IEEE Access, vol. 13, pp. 56283-56295, 2025, doi: 10.1109/ACCESS.2025.3554454.

[12] A. Amjad, L. -C. Tai and H. -T. Chang, "Utilizing Enhanced Particle Swarm Optimization for Feature Selection in Gender-Emotion Detection From English Speech Signals," in IEEE Access, vol. 12, pp. 189564-189573, 2024, doi: 10.1109/ACCESS.2024.3516790.

[13] E. Dikbiyik, O. Demir and B. Dogan, "BiMER: Design and Implementation of a Bimodal Emotion Recognition System Enhanced by Data Augmentation Techniques," in IEEE Access, vol. 13, pp. 64330-64352, 2025, doi: 10.1109/ACCESS.2025.3559339.

[14] L. Qu, T. Li, C. Weber, T. Pekarek-Rosin, F. Ren and S. Wermter, "Disentangling Prosody Representations With Unsupervised Speech Reconstruction," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 39-54, 2024, doi: 10.1109/TASLP.2023.3320864.

[15] C. Castorena, M. Cobos and F. J. Ferri, "An Incremental Selection Method for Semi-Supervised Speaker Adaptation in Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 32, pp. 2873-2877, 2025, doi: 10.1109/LSP.2025.3584714.

[16] D. -H. Cho, H. -S. Oh, S. -B. Kim and S. -W. Lee, "EmoSphere++: Emotion-Controllable Zero-Shot Text-to-Speech Via Emotion-Adaptive Spherical Vector," in IEEE Transactions on Affective Computing, vol. 16, no. 3, pp. 2365-2380, July-Sept. 2025, doi: 10.1109/TAFFC.2025.3561267.

[17] J. He, X. Shi, C. -H. Hu, J. Mi, X. Li and T. Toda, "M4SER: Multimodal, Multirepresentation, Multitask, and Multistrategy Learning for Speech Emotion Recognition," in IEEE Transactions on Audio, Speech and Language Processing, doi: 10.1109/TASLPRO.2025.3614428.

[18] D. -J. Min and D. -H. Kim, "Speech Emotion Recognition via Sparse Learning-Based Fusion Model," in IEEE Access, vol. 12, pp. 177219-177235, 2024, doi: 10.1109/ACCESS.2024.3506565.

[19] S. Akinpelu and S. Viriri, "Deep Learning Framework for Speech EmotionClassification:ASurveyoftheState-of-the-Art,"inIEEE Access, vol. 12, pp. 152152-152182, 2024, doi: 10.1109/ACCESS.2024.3474553.

[20] W. -C. Lin, K. Sridhar and C. Busso, "An Interpretable Deep Mutual Information Curriculum Metric for a Robust and Generalized Speech Emotion Recognition System," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 5117-5130, 2024, doi: 10.1109/TASLP.2024.3507562.