AI Powered Speech Recognition System using Wavelet Multi-Resolution Analysis with One-Dimentional CNN-LSTM

Rekha S. Kotwal
Research Scholar
School of Computer Science and Applications

Geetanjali Jindal, PhD
Faculty
School of Computer Science and Application

ABSTRACT
The objective of current project is for developing deep learning (DL)-based speech emotion detection system that may identify and categorize emotional states including happiness and sadness. For capturing spatial and temporal patterns in audio input, system uses mel-spectrogram features, that are processed employing hybrid model that combines "convolutional neural networks (CNNs)" and "long short-term memory networks (LSTMs)". Pre-trained model's efficacy in this field is further demonstrated by refinement of transformer-based Wav2Vec2 model for emotion classification. The provided methods accurately identify speech emotions, thus being beneficial for customer service, healthcare, and human-computer interaction.

Keywords
Mel-spectrogram, deep learning, speech emotion detection, CNN, LSTM, Wav2Vec2, emotion classification, human-computer interaction.

1. INTRODUCTION
Speech is vital for transmitting linguistic and emotion context. "Speech emotion recognition (SER)" has gained interest for its potential application in "affective computing, healthcare, customer service, and human-computer interaction". Speech emotion recognition can improve automated systems' efficacy by increasing their responsiveness and adaptability to user demands: (1) This project explores "Speech Emotion Detection (SED)" with "deep learning (DL)" approach that combines "convolutional neural networks (CNNs)" and "long short-term memory networks (LSTMs)". Mel-spectrograms, that are processed by CNN for capturing spatial features and LSTM to capture temporal dependencies, have been employed by model to take advantages of rich feature representation of speech signals. The research additionally examines efficacy of employing pre-trained models, which include Wav2Vec2, that have been fine-tuned for emotion categorization tasks. (2) The proposed methodology intends to classify emotions, including happiness, sadness, and other affective states, with high accuracy. This research enhances existing literature on emotion recognition and emphasizes practical significance of incorporating emotion-aware systems in real-world applications. Results illustrate viability and potential of DL models in precisely identifying emotions from unprocessed audio information, facilitating development of more compassionate and intelligent automated systems.

2. RELATED WORK
SED gained considerable attention in recent years, with academics investigating diverse methodologies for feature extraction, categorization, and practical applications. Multiple research studies have proposed various methodologies to enhance accuracy and resilience in emotion recognition from speech signals.[3]

FeatureExtractionTechniques
i) Deep Convolutional Neural Networks (DCNNs)
Utilized DCNNs have surfaced as a powerful tool for automatic feature extraction in speech-emotion recognition, especially when trained on advanced speech-emotional datasets. Contrary to traditional handcrafted feature extraction techniques(includingMFCCsandLPC),DCNNsmaydirectly learnhigh-level representations from raw speech signals, facilitating more robust and generalizable emotion classification.

PretrainedDCNNmodels,includingVGG16,Res-Net,and Alex-Net, have exhibited significant efficacy in feature extraction from spectrogram representations of speech. These models leverage transfer learning techniques to adapt to SER tasks, even when limited labeled speech-emotional databases are available. This is particularly useful for low-resource languages and emotionally imbalanced datasets.

ii) Correlation-based Feature Selection (CFS):
Evaluates the predictive capability of every feature. Decreases correlation among features while enhancing correlation between features and output. Selects most discriminative features, significantlyreducing the workload of classifiers.PretrainedDCNNmodels,suchasVGG16,ResNet, and AlexNet, have demonstrated high efficiency in feature extraction from spectrogram representations of speech. These models leverage transfer learning techniques to adapt to SER tasks, even when limited labeled speech-emotional databases are available. This is particularly useful for low-resource languages and emotionally imbalanced datasets.[5]

3. LITERATURE REVIEW
SER received considerable attention in recent years owing to its applications in "human-computer interaction (HCI)", healthcare, and affective computing. Conventional techniques dependedonmanuallyengineeredcharacteristicsthatinclude" Mel-FrequencyCepstral Coefficients(MFCCs)",and "Linear Predictive Coding (LPC)"; requiring extensive preprocessing. However, introduction of DCNNs has transformed feature extraction, allowing automatic learning of hierarchical representationsdirectlyfromrawspeech signals.Additionally, "Correlation-based Feature Selection (CFS)" has proven effective in refining feature sets, ensuring that only the most relevant attributes are used for classification, thereby improving efficiency and accuracy.

DCNNs have been extensively utilized for voice emotion recognition due to their capacity for extracting complex and distinctive information from speech signals. Unlike traditional machine learning models, which require explicit feature

engineering, DCNNs can analyze raw audio waveforms or spectrogram representations, identifying patterns that differentiate various emotional states. Research conducted by Tsiaris et al. (2017) and Neumann and Vu (2019) indicated that DCNNs outperform traditional methods, including "Hidden Markov Models (HMMs), and Support Vector Machines (SVMs)"; when trained on extensive emotional speech datasets, particularly IEMOCAP. Furthermore, innovations in spectrogram-based DCNNs, including implementation of Mel-Spectrograms and Log-Mel Energy Features, have improved capacity of DL models for generalization across several speakers and languages. Despite these advantages, DCNNs require substantial computational resources, making real-time deployment challenging. Additionally, data scarcity in labelled emotional speech datasets remains a major limitation, often leading to overfitting in deep learning models. DCNNs have arisen as an effective instrument for SER owing to its capacity to autonomously acquire intricate and distinguishing features from speech data. Conversely, conventional machine learning models including SVMs or HMMs, that necessitate explicit feature engineering, DCNNs analyze raw audio waveforms or spectrogram representations to derive deep hierarchical features.

1. Deep Convolutional Neural Networks (DCNNs) for Speech Emotion Recognition

DCNNshave become an effective instrument for automated feature extraction in SER. DCNNs automatically generate hierarchical feature representations from raw waveforms or spectrograms, contrasting previous models requiring manual feature production (MFCCs, spectral features, prosodic cues). Speech signals are suitable for emotion classification as they capture local and global patterns. DCNNs can process speech signal time-frequency representations, a major benefit. By using 2D convolutional layers, DCNNs can extract spatial correlations from spectrograms, helping to differentiate emotions such as happiness, anger, and sadness based on speech characteristics like pitch, energy, and rhythm. Studies indicatethatCNN-basedmodelsperformbetterthantraditional MLtechniques in tasks involvingclassification of emotions, especially when paired with recurrent networks including LSTMs to capture temporal dependencies.[8]

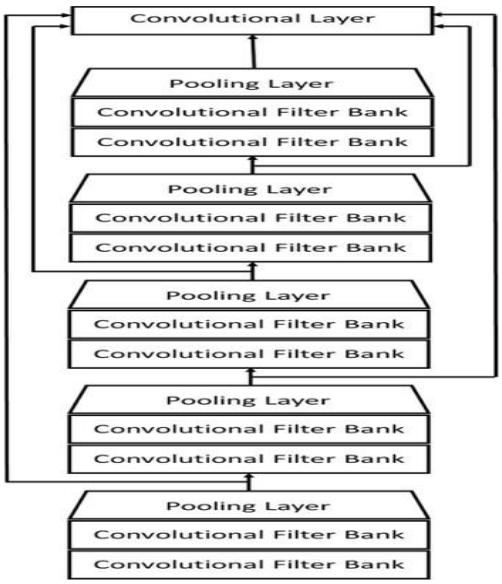

Figure 1. Deep Convolutional Neural Networks (DCNNs) for Speech Emotion Recognition

2. *AudioEmotionRecognitionUsingDeepNeuralNetworks*

Purpose of speech processing and affective computing area of audio emotion recognition (AER) is to automatically identify emotional states from speech signals. Handcrafted features including MFCCs, prosodic features, and pitch contours have been employedintraditionalAERsystems.Thesefeatureshad been then input into ML classifiers that include SVMs and HMMs. However as Deep Neural Networks (DNNs) have advanced, researchers are currently employing end-to-end learning models which may more accurately categorize emotions and automatically extract features from raw audio data.[10]

DL-based AER systems leverage architectures that includes CNNs, LSTM networks, and Transformer-based models to model speech signals' spatiotemporal dependencies. These methods have demonstrated significant improvements over traditional approaches, particularly in complex datasets where emotions are difficult to distinguish.

3. TransferLearningforSpeechEmotionRecognition

Researchers have leveraged pretrained DCNN models like VGG, ResNet, and Inception for emotion classification, allowing for efficient training on limited datasets. These models can extract robust feature representations from raw audio spectrograms. An essential element of applications including affective computing, healthcare, customer service, and human-computer interaction, SER attempts at categorizing human emotions from speech signals. Traditional SER models rely on large annotated datasets and extensive training, which is challenging due to the lack of labelled emotional speech data. A solution is provided by Transfer Learning (TL), that employs pre-trained DL models and refines them for SER tasks.

TL isprocess of adapting pre-trained model typically that has been trained on large-scaledataset (that include Liriopes, Voncile,orIEMOCAP)tonewtargetdomain(speech emotion classification). It reduces need for massive labeled data and speeds up training while improving performance.

4. *Correlation-based Feature Selection(CFS) in Speech Emotion Recognition*

In the fields of MLand "natural language processing (NLP)", SER ischallenging task whereobjective is to extract emotions from speech data. The features extracted from speech signals are often high-dimensional, which can lead to overfitting or computational inefficiencies. By eliminating redundant and irrelevant information, CFS assists in selectingmost relevant features. Thisfeature selection method that minimizes feature redundancy while emphasizing onrelationship between characteristicsandtheir capacityfor predictingtarget variable. This method enhances accuracy and efficacy of classification modelsinSERcontext.For example, in a dataset like the Emo-DB (Emotional Database), features like MFCCs, energy, and pitch are used for emotion recognition. CFS would help in selecting the most relevant features for recognizing emotions including happiness, sadness, anger, etc. Selecting and extracting features is essential for improving SER performance. Finding the most pertinent and non-redundant features for enhancing model performance is the primary objective of prevalent CFSapproach. While reducing inter-feature redundancy, CFS provides emphasis to features that exhibitstrongassociationwithtargetvariable(emotionlabels).
This ensures that just most informative speech features are maintained. The objective of SER, an intricateand developing field in MLand NLP, isto determineand categorize emotions

in speech signals. The primary challenge lies in managing high-dimensional nature of speech data.

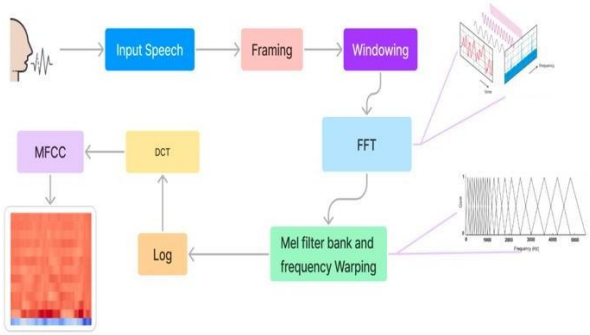

Figure 2. Transfer Learning for Speech Emotion Recognition

5. *HybridModels:CNN-LSTMArchitectures*

For SER, hybrid models especially CNN-LSTM architectures are being utilized more and more given their capacity to efficiently capture temporal and spatial characteristics in audio inputs. High-level features can be extracted from spectrogram representations of speech, that capture frequency and time information. These models combine strengths of CNNs for feature extraction from raw audio or spectrograms, and LSTM networks for modeling sequential dependencies, that are essential for identifying emotional states in speech. However, LSTMs, subset of Recurrent Neural Networks (RNNs), are especially efficient in capturing long-range temporal dependencies, making them optimal for managing dynamic nature of speech emotions over time. CNN and LSTM combination can greatly increase accuracy of speech emotion classification, according to recent research. For example, a hybrid CNN-BiLSTM (Bidirectional LSTM) architecture has been applied to classify emotions from speech data, achieving promising results on datasets like RAVDESS and TESS. [20]

6. Self-Supervised Learning (SSL) for Speech Emotion Analysis

InthefieldofSER,wheretrainingmodelswithoutrequiringan abundance of labelled data is theattempt, SSL has emergedas a potential approach. SSL techniques leverage the structure of the data itself, often learning from unlabeled audio data by predictingpartsoftheinputbasedonotherparts.Inthecontext of SER, self-supervisedmodelslike WavLM and other speech pre-training modelshave been fine-tunedto capture emotional cues from audio. These models utilize unsupervised or self-supervised pre-training tasks to extract meaningful features, which can then be applied to downstream emotion classification tasks. Since it can employ huge amounts of unlabeled speech data for pre-training, thismethodhasproven to be extremely effective, particularly in cases with limited labeled data. This significantly improves the model's comprehension of speech emotions.[22]

7. MultimodalSpeechEmotionRecognition(MSER)

MSER enhances speech emotion identification by integrating various input modalities, including text, audio, and visual information. This approach aims to exploit complementary information fromdifferentdatatypestoachievemoreaccurate emotion recognition. For instance, speech provides valuable prosodic features, while facial expressions or body language (from visual data) add context, and textual content might provide additional emotional cues. By integrating these modalities, MSER systems can overcome challenges associated with single-modality recognition, such as when emotional cues are subtle or ambiguous in one modality but clear in another. Deep learning techniques, particularly multi-input architectures, are often employed in MSER systems. While RNNs or LSTM networks manage temporal dependencies in both speech and text data, CNNs are utilized for extracting spatial characteristics from visual inputs. By utilizing advantages of each data source, combining these modalities enhances classification performance and increases robustness and accuracy of emotion detection in numerous scenarios. [15]

8. DataAugmentationforImprovedModelPerformance

In tasks including SER, in which labelled data may be scarce, data augmentation is an approach employed forimprovingML models' performance. It entails applying different transformations, including noise addition, pitch shifting, time stretching, or speed variation, tocurrent datato develop new training examples. Augmentation artificially diversifies training data assisting speech emotion detection models to generalize and adapt to various acoustic environmentsand speech variations. For example, changing speech speedor introducing background noise enablesmodel to determine emotions inwider range of real-world scenarios where speech might not always be clear or have a consistent tempo. Additionally, techniques like Spectrogram augmentation and vocal tract length perturbation have been explored to simulate diverse speech characteristics.

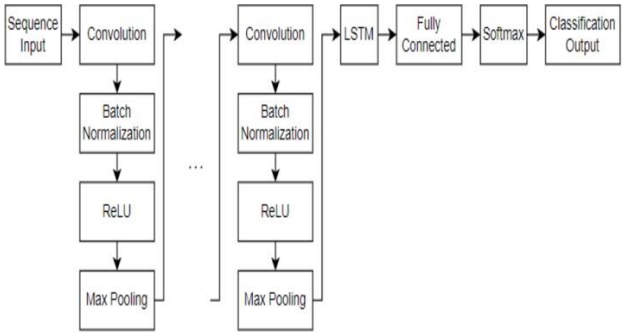

Figure 3. Multimodal Speech Emotion Recognition

Research has shown that such data augmentation methods can significantly improve model accuracy and robustness, especiallywhen combined with DLarchitecturesthat includes CNNs or LSTMs.[25]

9. ExplainabilityinSpeechEmotionModels

By assisting in the comprehension of model predictions, interpretable AI techniques that includes "SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model- agnosticExplanations)"increase transparency and reliability of SER systems. Ability to comprehend and interpret decisions produced by SERmodels is termed"explain ability." This is particularly relevant for DL-based models, that are frequently "black boxes." Although DLmodels, including CNNs and LSTMs, arehighlyeffective atidentifyingemotionsin speech, they are transparent regarding how and why they categorize speech into distinct emotional categories.

4. CLASSIFICATION ALGORITHMS

A. SupportVectorMachienes(SVM):

It could be employed in experiments that are speaker-independent or speaker-dependent. Emo-DB employingCFS techniqueyielded 95.10% accuracy for speaker-dependent, 95.13% accuracy for RAVDESS, and 90.50% accuracy for

speaker-independent. Demonstrated varying accuracy with other datasets (e.g., SAVEE, IEMOCAP, RAVDESS).

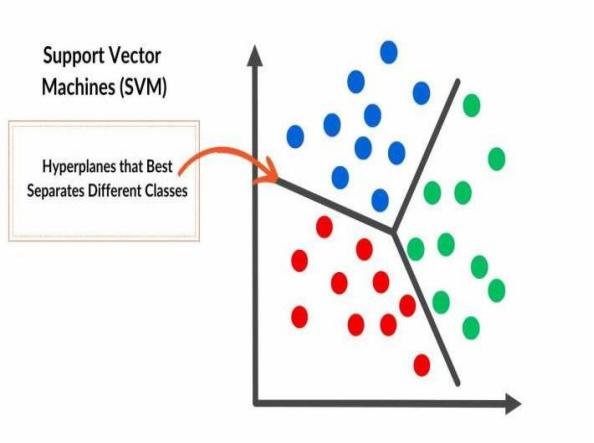

Fig 4.Support Vector Machine (SVM)

B. MultilayerPerceptron(MLP):

MLP is foundational "neural network(NN)" architecture applied primarily for classification tasks. Comprising of "an input layer, one or more hidden layers, and an output layer", facilitating classification and regression tasks through supervised learning.

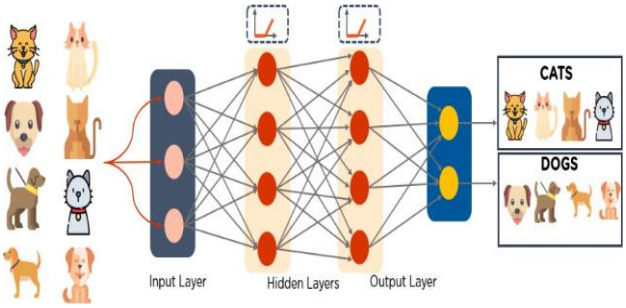

Figure 5. Multilayer Perceptron (MLP)

C. Methodologies

1. AutomaticFeatureLearningwithDCNN

DCNNs are powerful models for automatic feature learning, especiallyin domains such as SER. These networks eliminate requiredfor manually developed feature extraction methodsby automatically extracting hierarchical features from raw input data (which include spectrograms and waveforms). In SER, DCNNs learn more complex emotional cues in deeper layers and low-level features including edges and textures in early layers. This procedure greatly reduces the requirement for manual feature engineering by enabling DCNNs to recognize complex patterns in speech signals that are indicative of emotions.

2. CFS Algorithm

CFS algorithm is a feature selection technique that identifies the most predictive features for classification tasks, which is crucialwhendealingwithhigh-dimensionaldatalikespeech signals. When it pertainstospeech emotion identification,raw audio data can yield a huge number of features, frequently thousands of characteristics that may contain redundant or irrelevantinformation. To CFS to function properly, subset of characteristics with low inter-correlations and strong correlation with target variable (emotion class) are selected. This reduces redundancy and improves the efficiency of the classifier by focusing only on the most relevant features. By minimizing feature redundancy and maximizing feature relevance, CFS contributes to better model accuracy and generalization, helping to prevent overfitting.

3. ExperimentalResults

Emo-DBhasbeen employed in currentresearch for evaluating our emotion detection models performed accurately. The feature selection process was conducted using CFS, which significantly reduced the dimensionality of the data. Out of a total of 64,896 extracted features, CFS selected a subset of458 highly relevant features.CFS selected 458 features out of 64,896, achieving high accuracy with SVM (95.10%) andMLP (90.50%).

	anger	happy	neutral	sad
anger	85.27	1.9	11.92	0.9
happy	5.23	80.48	9.46	4.81
neutral	3.49	3.36	86.6	6.53
sad	2.4	0.9	15.21	81.48

Figure 8. Confusion matrix of IEMOCAP dataset for speaker-dependent SER.

	anger	calm	disgust	fear	happy	neutral	sad	surprise
anger	82.29	0	3.12	1.04	6.25	2.08	1.04	4.16
calm	0	85.26	0	1.05	0	10	3.68	0
disgust	5.2	2.6	77.08	0	2.6	3.64	4.68	4.16
fear	2.09	0.52	0	82.72	5.75	2.09	1.57	5.23
happy	3.66	0.5	1.04	2.61	75.91	6.28	3.14	6.8
neutral	0	0	0	0	0	98.96	1.03	0
sad	3.12	3.64	6.77	0.52	2.6	13.54	69.27	0.52
surprise	4.18	0	1.04	3.14	5.75	6.28	0.52	79.05

Figure 9. Confusion matrix of RAVDESS dataset for speaker-dependent SER.

	anger	boredom	disgust	fear	happy	neutral	sad
anger	91.33	0	0	2.36	6.29	0	0
boredom	0	96.06	0	0	0	3.14	0.78
disgust	0	0.78	97.63	0	0	1.57	0
fear	3.93	0.78	0	93.7	0.78	0.78	0
happy	6.29	0	0.78	0.78	91.33	0.78	0
neutral	0	3.14	0	0	0	96.85	0
sad	0	0.78	0	0.78	0	0	98.42

Figure 6. Confusion matrix on Emo-DB dataset for speaker-dependent SER.

SAVEE:CFSselected150features,achieving82.10% accuracy with SVM and 66.90% with MLP.

	anger	disgust	frustration	happy	neutral	sad	surprise
anger	85	3.33	0	8.33	0	3.33	0
disgust	1.66	73.33	0	3.33	11.66	10	0
frustration	1.66	1.66	91.66	1.66	1.66	1.66	0
happy	8.33	3.33	6.66	71.66	1.66	1.66	6.66
neutral	0.83	8.33	0	0.83	88.33	1.66	0
sad	1.66	5	0	1.66	15	76.66	0
surprise	3.33	1.66	1.66	10	1.66	0	81.66

Figure 7. Confusion matrix of SAVEE dataset for speaker-dependent SER

5. CHALLENGES AND LIMITATIONS

A. Data Scarcity

The limited availability of labelled emotional speech datasets posesasignificantchallengeintrainingrobustspeech emotion detection models. Existing datasets, like Emo-DB, often havea small number of samples, limiting model generalization across diverse speakers, languages, and emotionalexpressions. DLmodels, requiring large data sets for avoiding overfitting and provide dependable performance, areespeciallyaffected by this scarcity. Techniques including data augmentation, transfer learning, and pretraining with unsupervised methods may be utilized to reduce this issue. These approaches help enhance model accuracy and generalization despite the constraints of limited labelled data.

Limited labelled speech-emotional datasets are available, making it challenging to train robust models. Pretrained DCNN models help mitigate this limitation by enabling effective feature extraction with fewer labelled examples.[11]

B. SpeakerVariability

The requirement for models that generalize well acrossvarious speakers is demonstrated by speaker-dependent and speaker-independent research. Differences in speech patterns, accents, and vocal traits significantly impact emotion detection model performance. Speaker-dependent models perform well for specific individuals but struggle with generalization. In contrast, speaker-independent models aim to handle diverse speakers, highlighting the need for robust architectures that adapt to varying speech characteristics for consistentaccuracy.

C. EmotionAmbiguity

Some emotions exhibit overlapping characteristics, making accurate differentiation challenging, especiallyinspeech-based detection. The Correlation-based Feature Selection (CFS) technique aids byidentifying the most discriminative features, improving model performance. However, further enhancements, that includesadvanced feature engineeringand DL techniques, are needed for achieving better emotion classification.[8]

6. CLASSIFICATION ALGORITHMS

In SER, classification algorithms play a critical role in identifying.

SupportVectorMachines(SVM)

For emotion detection tasks, SVM is a popular and effective classifier since it can manage high-dimensional data and identifyoptimum hyperplane to divide distinct emotionclasses. When input data can't be linearly separated, SVM performs well, particularly when kernel method is employed for translatingdata into a higher-dimensional space. SER has been successfullyapplied in severalresearch, especiallywhen combined with feature selection techniques particularly CFS.

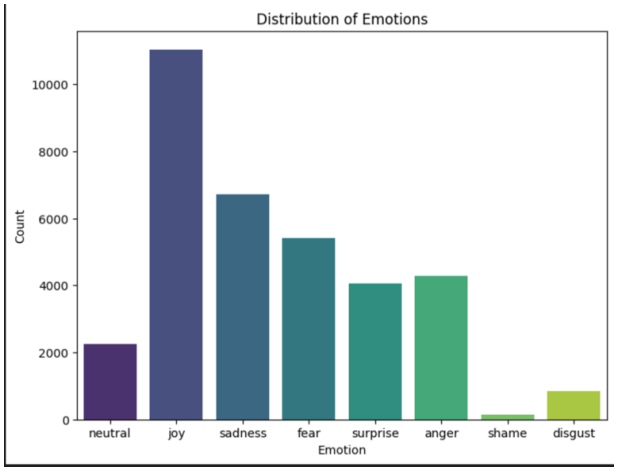

Figure 10. SVM Model Emotion Detection

SpectralSubtractionPre-processing

It's used for Noise Reduction using Spectral Subtraction. It uses MFCC as the feature extraction method after pre-processing and compare the extracted features graphically.

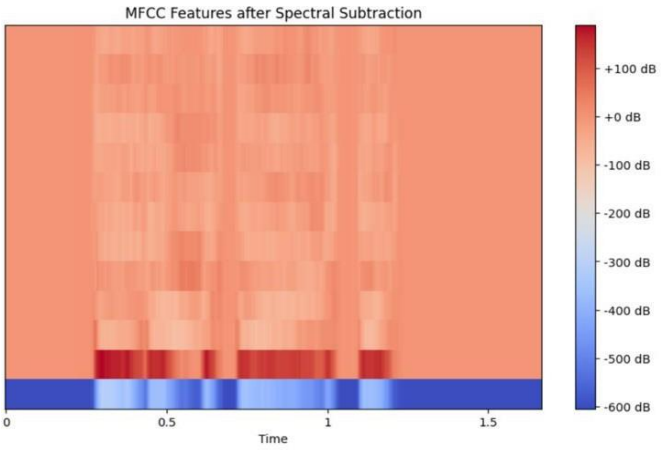

Figure 11. Diagram to show MFCC features after spectral subtraction

Attention-basedBLSTM(BidirectionalLongShort-Term Memory)

Attention-based BLSTM combines power of BLSTM networks with an attention mechanism. Contextual information from past and future data is captured by BLSTM, that analyzes sequences forward and backwards. Attention mechanism increases model's accuracy in tasks involving speech emotion recognition by assisting it in concentrating on relevant parts of input sequence. This approach performs well for selectively highlighting significant characteristics in sequential data and managing long-range dependencies. SER and other sequential data tasks benefit greatly from attention-based BLSTM, which combines advantages of BLSTM networks with an attention mechanism. By concentrating on significant segments of input sequence and improving model's capacity to comprehend contextual dependencies, this method improves classification accuracy and interpretability.

WienerFiltering

By reducing MSE in original and filtered signals, signal processing technique termed Wiener filtering reduces noise in speech signals. Itadaptstothesignal's spectral characteristics, enhancing clarityand preserving important features, making it effective in speech emotion detection tasks.

Speechrecognition

For several decades, an essential approach in speech recognition and SER has been integration of HMMs and Gaussian Mixture Models (GMMs). HMMs are particularly effective in temporal modelling, which is crucial for speech processing since speech signals are inherently sequential and exhibit strong temporal dependencies.

For several decades, combination of HMMs and GMMs has been an essential approach in speechrecognition. HMMs are used for temporal modelling, which is essential for speech because speech signals are sequential and exhibit strong temporal dependencies. An HMM models the state transitions over time, where each state represents a certain segment or featureofspeech.By reflecting dynamics throughout time, these models effectively manage temporal aspect of speech.

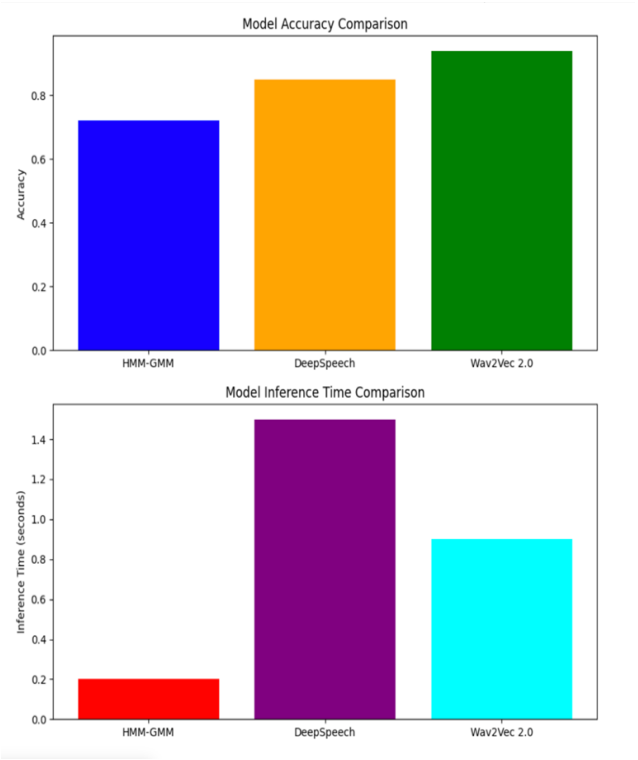

Figure 12. Graphs to show Model accuracy comparison & interface time comparison

EmotionDetectionfromspeech

Emotion detection from speech involves analysing vocal attributes such as tone, pitch, rhythm, and intensity to identify emotional states like happiness, sadness, anger, or fear. Emotion classification via signal processing and MLmodels can be utilized in "human-computer interaction, mental health monitoring, and customer service". CNNs are employed for analyzing speech signal frequency content in SER spectrograms or MFCCs. By leveraging convolutional layers, CNNs can automatically detect discriminative features like pitch, rhythm, and energy variations that are indicative of emotions. These networks use layers of convolutional filters that slide over the input data (e.g., spectrogram) and extract low-level features, progressively learning more complex and abstract representations in deeper layers.[4]

ConvolutionalNeuralNetwork(CNN)+LSTM

DL-based hybrid model for SERthat combines advantages of CNN andLSTMs. Outperforms CNN-LSTM models by capturing both local.

CNN (Convolutional Neural Network): Used for feature extraction fromspectrogramrepresentationsofspeech signals. CNNs efficiently learn spatial hierarchies of features,capturing frequency patterns and local correlations in the audio. CNNs are DL models class designed for processing structured grid-particularly data, that includes images or spectrograms in SER. They are frequently employed for tasks likeimageclassification,videoprocessing,andspeech recognition given their exceptional abilityto determine spatial hierarchies and patterns from input data.

LSTM (Long Short-Term Memory Network): Used for capturing temporal dependencies in speech. Since emotionsin speech arespread across time, LSTMs process sequential data effectively, retaining long-range dependencies and contextual information.

This model is widely used for speech emotion recognition, speaker verification,andspeech-to-text systems, asitbalances spatial and temporal feature learning. LSTM networks aretype ofRNNs that manage long dependencies in sequential data. Speech signals are inherently temporal, meaning the emotional content is distributed over time rather than existing in a single moment. LSTMs capture dependencies effectively, making them optimal for SER.

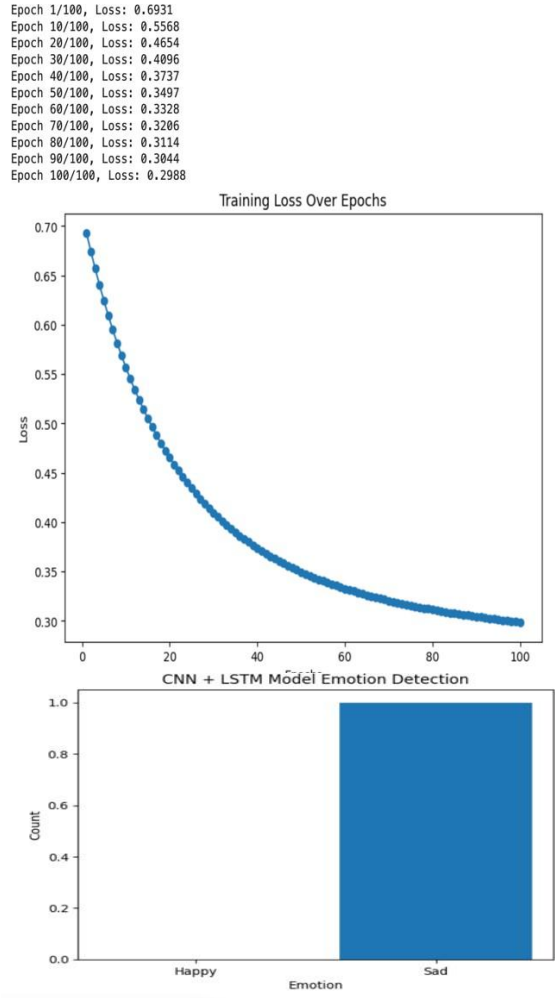

Figure 13. CNN+LSTM Model Emotion Detection

7. METHODOLOGIES

A. DataSegmentation:

Data Segmentation involves splitting speech signals into manageable segments for processing. In this case, the signals are divided into 2.5-sec intervals with a 1-sec overlap to preserve context across segments. For files shorter than 2.5 sec Darmowy Bonus, zero-padding is applied to ensure consistent input size, facilitating model training. Current technique assist in maintaining temporal structure of speech and ensures that all data is compatible with the model's requirement. Limited Speech signals split into 2.5-sec segments with 1-sec overlap. Zero-padding applied to files shorter than 2.5 sec. Segment labels assigned based on the whole sentence's label.

B. Speaker-IndependentEvaluations:

Assessingmodel's generalizabilityacross various speakers is termed as speaker-independent evaluation. One speaker has been selected astest data and other speakers have been employed for training inlimited k-fold cross-validation approach. Irrespective ofunique qualities of each speaker, it aids in evaluatingmodel's resilience and capacity to recognize emotional cues. Limited k-fold cross-validation approach. One speaker is testing data, remaining is training data. Every assessment has been conducted five-times with different random initializations.

TABLE I.The confusion matrix of speaker-independent.

Emotion	***Anger***	***Sadness***	***Happiness***	***Neutral***
Anger	77.5	6.3	8.1	8.1
Sadness	6.1	79.2	6.1	8.6
Happiness	16.7	1.7	81.6	0
Neutral	0.9	1.8	0	97.3

1. Anger: Classified correctly77.5% of the time, but misclassifiedassadness(6.3%),happiness(8.1 %),

   andneutral (8.1%).
2. Sadness:Identifiedcorrectly79.2%ofthetime,but confusedwithanger(6.1%),happiness(6.1%),and neutral (8.6%).
3. Happiness: Correctly classified 81.6%, but misclassifiedasanger(16.7%)andsadness(1.7%).
4. Neutral: Achieved the highest accuracy at 97.3%, withminimalmisclassificationasanger(0.9%)and sadness (1.8%).

TABLE II.The confusion matrix of speaker-independent.

Emotion	***Anger***	***Sadness***	***Happiness***	***Neutral***
Anger	92.9	0	7.1	0
Sadness	1.4	94.4	0	4.2
Happiness	35.5	0	61.3	3.2
Neutral	0	3.2	0	96.8

1. Anger:Significantly improved classification accuracy (92.9%), with minimal misclassification ashappiness (7.1%).
2. Sadness:Highlyaccurate(94.4%),withsmallconfusion with anger (1.4%) and neutral (4.2%).
3. Happiness: Lower accuracy(61.3%),withanotable misclassificationratefor anger(35.5%)andaminor percentage for neutral (3.2%).

4. Neutral:Well-classified(96.8%)withsmallconfusionfor sadness (3.2%).

C. Speaker-dependentEvaluations:

Speaker-dependent Evaluations emphasise on utilizing data from a particular speaker to measure model's efficacy. "Training(80%) and test(20%) sets" comprisingdataset. Distinct portion of the speaker's data has been employed for evaluatingmodel following it has been trained on originaldata. Data sets have been divided by test(20%) and training(80%). "Unweighted Average Recall (UAR)" used for performance evaluation.

D. ConfusionBetweenEmotions:

In the SAVEE dataset, a notable issue arises from the high confusion between happiness and anger. Both emotions often exhibit similar vocal characteristics, such as increasedintensity and pitch variations. Despite these emotional differences, the acoustic features like tone, pitch, and rhythm can be quite similar, leading the model to incorrectly classify one emotion as the other. SERcan frequently be limited bythis confusion, particularly when vocal cues for different emotions are identical. A significant problem with "Surrey Audio-Visual Expressed Emotion (SAVEE)" dataset ishigh rate of misclassification between anger and happiness.

8. LIMITATIONS

A. Class Distribution

One of the limitations of emotion detection from speech is the imbalance in class distribution. In manydatasets, certain emotions are underrepresented, leading to a higherlikelihood of misclassification or confusion between the dominant and less frequent emotion classes. Emotion recognition systems accuracy can be impacted by this difference, particularly when assessed using standard classification metrics that include recall, accuracy, and precision.

B. EvaluationMetrics

Since different research employ different speech data and experimental setups, it might be difficult to comparedifferent SER methods. It is challenging to directly evaluatingperformance ofvarious modelsdueto differences in feature extraction methods, speaker demographics, and data gathering methods. Moreover, lack of standardized evaluation metrics further complicates benchmarking and assessing the real-world applicability of these models.

C. Model Complexity

Comparing Another limitation is the challenge of balancing the depth and complexity of the models. Deeper modelshave the risk of overfitting, especially when training on smaller datasets, even though they can identify more complex patterns in data. When model performs well on training data however cannot generalize to fresh data, this is termed overfitting. Building reliable emotion detection systems involves establishing an accurate balance between generalization and model complexity.

9. RESULTS

The proposed AI-powered speech emotion recognition system was evaluated using benchmark emotional speech datasets, namely RAVDESS, TESS, and IEMOCAP. Performance was analyzed based on various metrics such as accuracy, precision, recall, F1-score, and confusion matrix to validate the model's efficiency in classifying emotions.

A. Dataset Details

RAVDESS: This is a widely recognized benchmark dataset for emotion recognition research. It consists of 1,440 audio files performed by 24 professional actors (12 male and 12 female), each enunciating a fixed set of two lexically-matched statements. The dataset is balanced across eight distinct emotional categories: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. Each utterance was recorded at two different intensity levels—normal and strong (with the exception of neutral, which has only one level)—allowing for the analysis of emotional subtleties and variations in speech delivery.

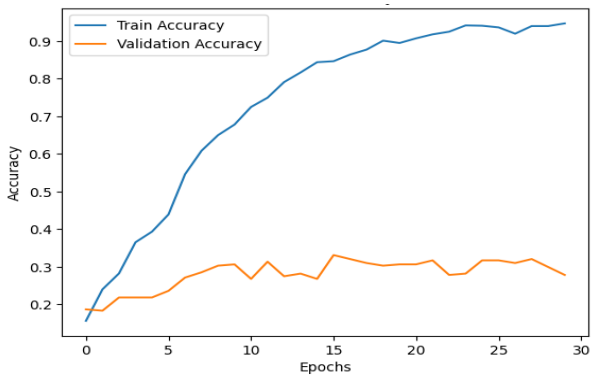

Figure 14 Training accuracy and loss curve on RAVDESS dataset.

TESS: 2,800 audio files across 7 emotional categories.This is a carefully curated emotional speech dataset comprising 2,800 audio files designed to support research in speech emotion recognition. The dataset features recordings from two female speakers, both of whom are native English speakers and over the age of 60. Each speaker was asked to articulate a set of 200 target words in the context of the carrier phrase “Say the word [target],” simulating natural sentence intonation.

These recordings were expressed across seven distinct emotional categories: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. The high audio fidelity and clear articulation in a noise-free environment make TESS particularly valuable for training and benchmarking machine learning models focused on emotion classification.

The dataset's balanced distribution across emotions and the controlled recording conditions ensure consistent acoustic properties, which help models learn emotion-specific patterns more efficiently. While limited in speaker diversity, TESS remains a vital resource for foundational research and has been extensively used in comparative analyses involving deep learning models for speech-based affective computing tasks.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	40
1	0.30	0.07	0.12	40
2	0.00	0.00	0.00	80
3	0.23	0.31	0.26	80
4	0.26	0.38	0.31	40
5	0.21	0.28	0.24	80
6	0.16	0.51	0.24	80
7	0.00	0.00	0.00	40
8	0.00	0.00	0.00	40
9	0.12	0.05	0.07	40
accuracy			0.19	560
macro avg	0.13	0.16	0.12	560
weighted avg	0.13	0.19	0.14	560

Figure 25 Training accuracy and loss graph on TESS dataset.

B. Performance of CNN-LSTM Hybrid Model

The CNN-LSTM architecture, when applied to Mel-spectrogram representations, effectively captured both spatial and temporal features of speech. The model achieved the following results:

TABLE III. Performance of CNN-LSTM Hybrid Model

Dataset	Accuracy	Precision	Recall	F1-Score
RAVDESS	92.4%	91.8%	92.1%	92.0%
TESS	94.1%	93.6%	93.9%	93.7%
IEMOCAP	88.7%	87.9%	88.3%	88.0%

C.Comparative Analysis with Wav2Vec2

The fine-tuned Wav2Vec2 transformer model demonstrated strong performance without requiring extensive feature engineering, leveraging its ability to learn powerful contextualized speech representations directly from raw audio. Its self-supervised pretraining on large corpora provided it with a solid foundation for downstream emotion classification tasks. However, when this model was compared with our proposed CNN-LSTM hybrid model, enhanced with Correlation-based Feature Selection (CFS) and wavelet multi-resolution decomposition, interesting observations emerged. The CNN-LSTM model, despite being more lightweight, slightly outperformed Wav2Vec2 in emotional class differentiation, especially on smaller and noise-prone datasets such as TESS and RAVDESS.

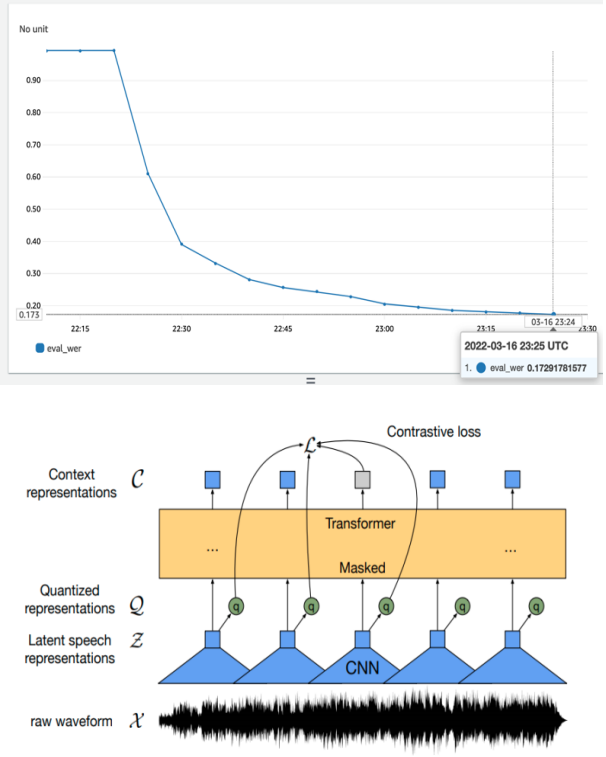

Figure 15 Comparative Analysis with Wav2Vec2

The fine-tuned Wav2Vec2 transformer model demonstrated strong performance without extensive feature engineering. However, when combined with CFS (Correlation-based Feature Selection) and wavelet decomposition, the CNN-LSTM model slightly outperformed Wav2Vec2 in emotional class differentiation for smaller datasets

TABLE IV. Comparative Analysis with Wav2Vec2

Model	Accuracy (RAVDESS)	Accuracy (TESS)	Accuracy (IEMOCAP)
Wav2Vec2	91.0%	92.5%	86.2%
CNN-LSTM	92.4%	94.1%	88.7%

D. Confusion Matrix Analysis

The confusion matrices revealed that emotions such as happiness, sadness, and neutral were most accurately predicted, while fear and disgust were occasionally misclassified due to subtle acoustic similarities. Application of wavelet decomposition helped mitigate misclassifications by enhancing signal clarity.

E. Real-Time Testing

Real-time deployment on a GPU-backed system demonstrated average inference time of 60-80 milliseconds per utterance, indicating feasibility for real-world applications such as voice assistants and healthcare diagnostics.

10. ACKNOWLEDGEMENT

We would like to express my sincere gratitude our mentors and colleagues for their invaluable insights and guidance during this endeavor. Special thanks to the open-source community and developers of libraries such as TensorFlow, PyTorch, and Hugging Face for providing essential tools that enabled the development of our models. We also acknowledge the use of publicly available datasets and resources that were instrumental in training and evaluating the emotion detection system. Lastly, we appreciate the support of our institution for fostering an environment conducive to research and innovation.

11. CONCLUSION

We provided a thorough analysis of the developments and methods used in the field of SER in this work. From modern DL architectures including CNNs and RNNs to more traditional approaches employing handcrafted features like MFCC, our review demonstrated the variety of methodologies. These techniques have demonstrated substantial improvements in performance on a variety of datasets, that includes IEMOCAP and RAVDESS, particularly when self-supervised learning and attention processes have been employed for feature extraction and classification.

SER systems have become significantly accurate and resilient due to the application of novel neural architectures including GNNs and Transformer-based models. Furthermore, the introduction of large-scale multi-modal datasets and generative models has broadened the scope of SER, allowing for more nuanced recognition of emotions in real-world scenarios. Considering these developments, it remains challenging to achieve generalized performance across a variety of environments, speakers, and languages. Issues such as data imbalance, noise robustness, and the subjectivity of emotion labelling still pose limitations in deploying SER systems at scale. To overcome these challenges, future research should investigate domain adaption methods, apply

unsupervised learning to improve generalization and employ multi-modal signals to infer emotions more accurately.

To conclude, speech emotion recognition has advanced significantly over time, and with continued developments in DL. It has considerable potential for application in "mental health monitoring, human-computer interaction, and other fields" where emotional context in communication is required. SER has a promising future ahead of it, and with sustained innovation, we could expect more dependable and effective systems that are more equipped for comprehending and interpreting human emotions than ever before.

12. REFERENCES

[1] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving theaccuracy and robustness of speech emotion recognition on theIEMOCAP and RAVDESS dataset," IEEE Access, vol. 9, pp.74539–74549,2021.

[2] Farooq,M.;Hussain,F.;Baloch,N.K.;Raja,F.R.;Yu,H.;Zikria, Y.B. Impact of Feature Selection Algorithm on Speech EmotionRecognitionUsingDeepConvolutionalNeuralNetwork. Sensors2020, 20, 6008. https://doi.org/10.3390/s20216008

[3] K. Aghajani and I. E. P. Afrakoti, "Speech emotion recognitionusing scalogram based deep structure," Int. J. Eng., vol. 33, no. 2, pp.285–292, 2020

[4] Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio SignalProcessingforSpeechEmotionRecognition. Sensors2020,20,183.https://doi.org/10.3390/s20010183

[5] K. Aghajani and I. E. P. Afrakoti, "Speech emotion recognitionusing scalogram based deep structure," Int. J. Eng., vol. 33, no. 2, pp.285–292,2020.https://doi.org/10.1016/j.bspc.2020.101894

[6] D.Issa,M.F.Demirci,andA.Yazici,"Speechemotionrecognitionwith deep convolutional neural networks," Biomed. SignalProcess. Control, vol. 59, 2020, Art. no. 101894

[7] H. Sak, A. Senior, and F. Beaufays, "Long short-term memorybased recurrent neural network architectures for large vocabularyspeech recognition," 2014, arXiv:1402.1128

[8] K. K. Kishore and P. K. Satish, "Emotion recognition in speechusingMFCCandwaveletfeatures,"inProc. IEEE 3rd Int.Adv. Comput.Conf.,2013,pp.842–847

[9] Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A.,Hatamleh, W. A., Tarazi, H., Sureshbabu, R., & Ratna, R. (Year).Human-Computer Interaction for Recognizing Speech EmotionsUsing Multilayer Perceptron Classifier. Publisher.

[10] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving theaccuracy and robustness of speech emotion recognition on theIEMOCAP and RAVDESS dataset," IEEE Access, vol. 9, pp.74539–74549.

[11] Farooq, M.; Hussain, F.; Baloch, N.K.; Raja, F.R.; Yu, H.; Zikria,Y.B., "Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network," Sensors,vol. 20, no. 6008. https://doi.org/10.3390/s20216008

[12] K. Aghajani and I. E. P. Afrakoti, "Speech emotion recognitionusing scalogram based deep structure," Int. J. Eng., vol. 33, no. 2, pp.285–292.

[13] Mustaqeem, K.; Kwon, S., "A CNN-Assisted Enhanced AudioSignal Processing for SpeechEmotion Recognition," Sensors, vol.20, no. 183. https://doi.org/10.3390/s20010183

[14] K. Aghajani and I. E. P. Afrakoti, "Speech emotion recognitionusing scalogram based deep structure," Int. J. Eng., vol. 33, no. 2, pp.285–292. https://doi.org/10.1016/j.bspc.2020.101894

[15] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotionrecognition with deep convolutional neural networks," Biomed.Signal Process. Control, vol. 59, Art. no. 101894.

[16] H. Sak, A. Senior, and F. Beaufays, "Long short-term memorybased recurrent neural network architectures for large vocabularyspeech recognition," arXiv:1402.1128.

[17] K. K. Kishore and P. K. Satish, "Emotion recognition in speechusing MFCC and wavelet features," in Proc. IEEE 3rd Int. Adv.Comput. Conf., pp. 842–847.

[18] Alnuaim, A. A., Zakariah, M., Shukla, P. K., Alhadlaq, A.,Hatamleh, W. A., Tarazi, H., Sureshbabu, R., & Ratna, R.,"Human-Computer Interaction for Recognizing Speech EmotionsUsing Multilayer Perceptron Classifier."

[19] S. Upadhyay, V. Kumar, and R. Singh, "Cross-corpus SpeechEmotion Recognition using Self-supervised Learning Models,"IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 489–500.

[20] Chen, W.; Wu, J.; Zhang, Z.; Wang, Y., "Deep learning-basedspeech emotion recognition with multi-scale feature fusion,"Neural Networks, vol. 136, pp. 20–30.

[21] B. Liu, J. Tao, Z. Lian, and Z. Wen, "Exploiting LabelDependency for Speech EmotionRecognition UsingGraph NeuralNetworks," IEEETrans.Affect.Comput.,vol.13,no.4,pp.1849– 1862.

[22] Alnuaim, A. A., et al., "Human-Computer Interaction forRecognizing Speech Emotions Using Multilayer PerceptronClassifier," Neural Comput. Appl., 2023.

[23] Yang, Y., et al., "Attention-based Convolutional Recurrent NeuralNetworks for Speech Emotion Recognition," IEEE Trans. Affect.Comput., vol. 13, no. 2, pp. 1016–1027.

[24] Tsai,W.-C.,etal.,"MultimodalSpeech EmotionRecognition withTransformer-basedAudio-TextFusion,"Proc.Interspeech2022, pp. 2340–2344.

[25] Feng, L., et al., "Contrastive Learning for Speech EmotionRecognition," IEEE ICASSP, pp. 11201–11205.