International Journal of Applied and Advanced Multidisciplinary Research (IJAAMR)
Vol. 1, No. 1, 2023 : 25-46

Hybrid Acoustic and Deep Learning Approach for Enhanced Speech Emotion Recognition

Mrs. Meenakshi Thalor, Gargi Bharshankar
Aissms Ioit, India
Corresponding Author : Gargi Bharshankar gargibshankar@gmail.com

A R T I C L E I N F O

*Keywords:*Emotion recognition, Human-computer interaction,Hybrid architecture,Deep networks,Classification accuracy

Received : 08, July
Revised : 18, August
Accepted: 28, September

A B S T R A C T

Emotion recognition in speech is a key research topic in human-computer interaction. Understanding emotions in conversations can shed light on a person's well-being. This study introduces a hybrid architecture that combines acoustic and deep features for improved speech emotion recognition. Acoustic features like RMS energy and MFCC are extracted from voice records. Additionally, sound spectrogram images are processed using deep networks like VGG16 and ResNet to obtain deep features. These are merged into a hybrid feature vector, refined by the ReliefF algorithm. For classification, the Support Vector Machine is employed. Testing on datasets like RAVDESS and EMO-DB yielded accuracy rates up to 90.21%. Our method consistently outperformed existing techniques in accuracy.

INTRODUCTION

Speaking is the fundamental mode of human communication, known for its speed and efficiency. When we speak, air travels from our lungs through the trachea to the larynx, where it vibrates the vocal cords, producing speech signals. In recent years, there has been a growing interest in research related to speech emotion recognition. Automatically recognizing and quantifying human emotions has become a cutting-edge area of study spanning fields from biomedical engineering and psychophysiology to computer engineering and artificial intelligence.

Emotions, often accompanied by meaningful attitudes, are potent mental activities that can be observed through various bodily expressions, including speech, facial gestures, and body movements, all integral to a person's emotional state. Voice signals not only convey the speaker's identity but also transmit their emotional state to others. With the rise in demand for intelligent systems and the increased processing capabilities of computers, emotion recognition has gained prominence in human-computer interaction. Autonomous speech emotion recognition systems simulate human emotions through computer algorithms, using features such as accentuation, intonation, and pauses, often employing spectrum-based features to match them with target emotions. These systems generally consist of three stages: speech data preprocessing, emotion feature extraction, and emotion classification.

During speech, emotions can exhibit diversity and variation due to factors like physical conditions and external surroundings. Consequently, the development of robust categorization frameworks and emotion features that encapsulate essential knowledge is critical for emotion recognition. Speech emotion recognition presents a significant challenge for researchers, given its wide-ranging applications in areas such as voice surveillance, e-learning, clinical studies, lie detection, entertainment, computer games, and call centers. However, the subjectivity of emotions poses difficulties, as different individuals may perceive the same emotions differently, leading to uncertainties in defining basic emotion classes. Additionally, there is considerable ambiguity surrounding the selection of appropriate emotional features, with no predefined feature set designated for emotion recognition.

Furthermore, the presence of background noise in audio recordings, stemming from real-world sounds, can substantially impact the performance of machine learning models. Traditional approaches to speech emotion recognition involve extracting features that represent the acoustic content of speech, followed by the application of various machine learning techniques to establish relationships between these features and predetermined emotion labels. Commonly used methods include Support Vector Machines (SVMs), Hidden Markov Models (HMMs), and neural networks. While SVMs offer relatively good predictions with minimal effort, constructing and training neural networks and HMMs can be labor-intensive, requiring substantial computational resources and time. Today, deep learning models are being employed for tasks such as face recognition, voice recognition for the Internet of Things, and speech emotion recognition. A significant advantage of deep

learning techniques lies in their ability to automatically select relevant features from audio files associated with specific emotions in speech emotion recognition tasks.

LITERATURE REVIEW

Table 1. relevant research

Sr. No.	Paper Name and Author	Journal Name and Year	Short Introduction	Limitations
1.	"A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features" by MEHMET BİLAL ER	Date of publication December 7, 2020, in IEEE Access	This study innovatively combines acoustic and deep learning features using pre-trained neural networks to enhance speech emotion recognition, aiming for robust real-world applicability and setting new efficiency benchmarks.	This study may face challenges with the direct application of image-optimized networks to speech, computational demands of complex models, the inherent subjectivity of emotions, potential overfitting from combined features, and real-world noise interference, all while ensuring its methods remain contemporary in benchmark comparisons.
Sr. No.	Paper Name and Author	Journal Name and Year	Short Introduction	Limitations
2.	Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings by Patrick Schlegel, Stefan Kniesburges, Stephan Dürr, Anne Schützenberger & Michael Döllinger	in Scientific research at Nature in 2020	The study used High-Speed Video (HSV) to record 358 individuals (260 females, 98 males) phonating the vowel /i/. Subjects were categorized into healthy or those with Functional Dysphonia (FD), based on clinician diagnosis, with recordings approved by Friedrich-Alexander-University Erlangen-Nürnberg's ethics committee.	Only one vowel (/i/) was analyzed. Diagnosis might be subjective based on individual clinicians. Just one recording per subject might not capture voice variances.
Sr. No.	Paper Name and Author	Journal Name and Year	Short Introduction	Limitations
3.	Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability by Mimma Nardelli, Gaetano Valenza, Alberto Greco, Antonio Lanata	in IEEE Access in 2015	The study analyzed emotional responses to affective sounds using Heart Rate Variability (HRV) from ECGs in 27 volunteers, achieving approximately 84.5% accuracy in classifying arousal and valence emotional dimensions.	Emotional detection relied solely on HRV data, omitting other potential physiological indicators.

Sr. No.	Paper Name and Author	Journal Name and Year	Short Introduction	Limitations
4.	Bi-modal emotion recognition from expressive face and body gesturesby Hatice Gunes	in Journal of Network and Computer Applications in 2007	This study explores emotion recognition by integrating both facial and upper-body gestures from video sequences. Using two cameras, the combined approach outperforms methods focusing solely on face or body.	The manual selection of expressive frames may introduce bias, and relying on dual cameras could complicate practical implementations.

Sr. No.	Paper Name and Author	Journal Name and Year	Short Introduction	Limitations
5.	Speech emotion recognition based on long short-term memory and convolutional neural networks by LU Guanming;YUAN Liang;YANG Wenjuan;YAN Jingjie;LI Haibo	in Journal of Nanjing University of Posts and Telecommunications(Natural Science Edition). 2018	The study proposes a combined LSTM and CNN approach for speech emotion recognition, which was tested on three databases. This combined method outperformed traditional techniques and standalone LSTM or CNN methods.	Performance varied across databases, indicating potential sensitivity to dataset specifics.

METHODOLOGY

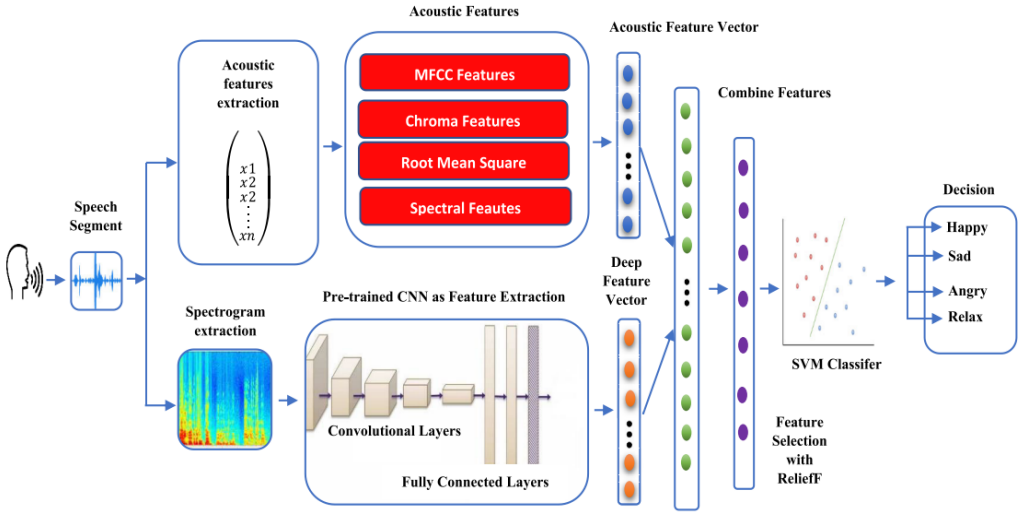


FIGURE 1. Proposed Method.

This study introduces a method combining acoustic features, deep features, and a pre-trained CNN-SVM model to enhance speech emotion recognition. While many studies use acoustic and deep features separately, this

study merges them to capture a richer emotional context in speech. Acoustic features alone might miss certain emotion patterns in speech and are based on subjective interpretations. On the other hand, deep features from networks like VGG16, ResNet18, ResNet50, and others offer a broader insight. Pre-trained CNNs, especially those trained on the ImageNet dataset, are used for feature extraction. By fusing acoustic and deep features, a hybrid feature vector is created. The ReliefF algorithm then refines this vector, selecting the most impactful features, which are subsequently used to train the SVM classifier. The model follows these steps:

1. Extract acoustic features from cleaned sound signals.
2. Obtain spectrogram images from signals and augment data.
3. Utilize CNNs trained on ImageNet for feature extraction.
4. Merge the deep and acoustic features into a hybrid vector.
5. Use ReliefF for feature refinement.
6. Train the SVM classifier with the refined hybrid feature vector.

A. ACOUSTIC FEATURE EXTRACTION

Acoustic analysis aims to deconstruct the speech signal into its constituent elements and offer a parametric evaluation of them. Acoustic features encompass physical attributes like frequency, loudness, and amplitude. Prior to extracting these features, speech signals undergo pre-processing to eliminate extraneous information, such as noise caused by surrounding conditions. For noise reduction, this research employs the Butterworth filter. Additionally, speech signals are segmented into 30ms frames, with a 15ms overlap.

The LibROSA toolbox serves as the tool of choice to derive acoustic features from diverse categories. Recognized as a popular library for analyzing music and sound, LibROSA facilitates the extraction of several acoustic attributes. These include Root Mean Square energy (RMS), MFCC, Chroma, Spectrum centroid, Spectral entropy, Skewness, Attack time, and Zero crossing rate, cumulating in an acoustic property vector of size 32. The RMS, a gauge of an audio signal's loudness, is computed by taking the square root of the average of the squared sound sample amplitudes. The mathematical representation of RMS is provided in Equation 1

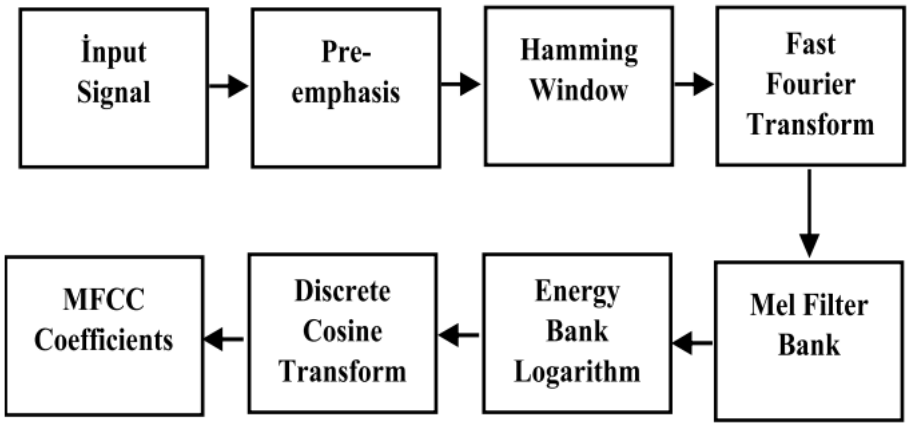
$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}} \qquad (1)$$

Spectrum Centroid: Often related to a sound's brightness, the spectrum centroid determines the center of mass of the spectrum. Higher centroid values are indicative of higher frequencies.

Spectral Entropy: This value takes into account the probabilities of the power spectrum components of the signal. It evaluates the normalized power distribution in the frequency domain as a probability distribution.

Skewness: This represents the asymmetry level of a distribution around its mean. Specifically, it refers to the average skewness coefficient of the spectral distribution in the lower frequency bands.

MFCC (Mel-Frequency Cepstral Coefficients): Rooted in human auditory perception, MFCC is a popular feature extraction method in sound processing. It aims to capture unique speaker values by mimicking the frequency selectivity of the human ear. The steps for extracting MFCC features are detailed in a subsequent figure.



The conversion between the Mel scale (M) and the frequency scale (Hz) can be accomplished using the specified equations (2 and 3).

$$m = 295 log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

$$f = 700\left(10^{\frac{m}{295}} - 1\right) \quad (3)$$

The perceptual structure can be represented using a triangular band-pass mel filter bank. After applying this filter bank, the Mel Frequency Cepstrum coefficients (MFCC) are derived by conducting a discrete cosine transform. The exact formula for calculating the MFCC is denoted as equation 4.

$$MFCC_i = \sum_{k=1}^{20} X_k . Cos\left[i.\left(\frac{k-1}{2}\right).\frac{\pi}{20}\right] \quad i = 1, 2, \ldots M \quad (4)$$

Attack Time: This measures the duration it takes for a signal to reach its peak. Essentially, it predicts the timeframe over which the amplitude of a signal ascends to its maximum value.

Chroma: Chroma pertains to the energy density around musical notes and offers crucial insights into a sound's harmonic content. Western music features 7 notes, and considering the subdivisions between most of these notes (excluding between E and F), a total of 12 features can be derived by accounting for the intermediate sounds.

Zero-crossing Rate: This indicates how frequently a signal transitions across the zero line, essentially signifying a change in its sign. Represented on the X-axis, it reveals the number of times the signal has crossed this line. This can be an indicator of both the frequency of a sound and its noise levels.

B. SPECTROGRAM EXTRACTION

During conversations, there can be moments of silence devoid of any emotion. Such segments can complicate emotion recognition. Yet, given the relatively short duration of the speech samples in the datasets, the entirety of these recordings was considered for spectrogram extraction. Since speech signals vary over time, it's essential to process them in short frames.

For spectrogram extraction, these signals are divided into frames of 30 ms each, with each frame overlapping half of its predecessor. Post framing, windowing is performed on the signal to avert potential discontinuities at frame boundaries. The study employs the widely-used "Hamming window" technique. The Hamming Window is designed to minimize undesirable side effects at frame extremities, and it ensures that the signal is aptly prepped for Fourier transform. The methodology involves multiplying the framed sound signal with the hamming window to produce the final windowed signal. The formula for the Hamming window is detailed as equation 5.

$$w[n] = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right), \quad N : Windowlength \tag{5}$$

Following the application of the Hamming window, the signal undergoes Fourier transform, converting it from the time domain to the frequency domain. Specifically, the study employs the Fast Fourier Transform (FFT) technique. In this process, each frame, composed of N samples, is subjected to FFT, effecting the transformation from the time domain to the frequency domain. A set comprising N samples is defined according to equation 6. In the final stage, the power spectrograms of the signals that have undergone Fourier transform are extracted. An illustrative example, showcasing a speech audio signal and its corresponding spectrogram image, is provided in Figure 3.

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, \quad n = 0, 1, 2, \ldots\ldots, N-1 \tag{6}$$

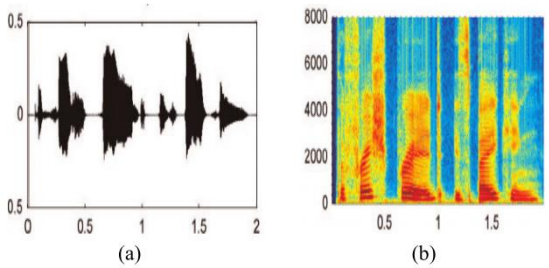


FIGURE 3. **Illustration of Speech Sound Signal and Spectrogram (a) Speech Sound Signal, (b) Spectrogram.**

C. DATA AUGMENTATION

When there's a scarcity of original data, data augmentation emerges as a solution. It involves generating supplementary training data samples by making systematic modifications to the existing data in the training set. A fundamental principle in data augmentation is that the labels of newly formed data, derived from alterations to labeled data, remain unchanged. Various augmentation techniques exist, such as image rotation, horizontal and vertical flipping, noise addition, and color manipulation. For this study, two specific procedures were employed on the signal prior to extracting the spectrogram:

- **Background Noise:** Random noise, ranging between 0.1 and 0.5, was added to the audio samples.
- **Time Shifting:** Audio samples were offset from their starting point by up to 0.3 seconds, ensuring the original sample duration remained unchanged. Following these augmentations, the newly derived audio samples were incorporated into the original dataset to bolster the training examples. The conceptual outline of this data augmentation process is encapsulated in Algorithm 1.

Algorithm 1: Data Augmentation Algorithm

Input:

1. sp: Spectrogram
2. Sr: Sample Rate
3. bg: Background Noise
4. tm: Time Shifting,
5. bgnr: Background Noise Range
6. tsh: Time shifting rate

Algorithm:

1. Initialize and assign input parameters (sp, sr, bg, tm, bgnr, tsh)
2. sp = read_folder('foldername/'+Filename');
3. for i: = 1 to length(sp) do
4. data = Read Spectrogram File(sp.Name);
5. bg = add.random.bgnoise(data, bgnr);
6. tm = Time.shifting(data, sr ∗ tsh)
7. Write spectrogram(bg_noise.jpg, bg);

8. Write spectrogram(time_shifting.jpg, tm);
9. End for training examples.

D. DEEP FEATURE EXTRACTION FROM PRE-TRAINED CNN MODELS

Pre-trained Convolutional Neural Networks (CNNs) predominantly consist of three layer types: convolution, pooling, and fully connected layers. The convolution and pooling layers focus on feature extraction, whereas the fully connected layers channel these extracted features towards the final output for the purpose of classification. In the convolution layer, filters of specified dimensions traverse the input image from left to right. The convolution process is represented by a formula, with 'M' denoting the feature map presented in the equation, while 'w' stands for the convolution kernel of size (x, y).

Convolution formula is given in equation 7.

$$M(i,j) = \left(R^{*}w\right)(i,j) = \sum_{x}\sum_{y} R(i-x, j-y)\, w(x,y) \quad (7)$$

E. DEEP FEATURE EXTRACTION FROM PRE-TRAINED CNN MODELS

Pre-trained CNNs primarily comprise layers such as convolution, pooling, and fully connected layers. The pooling layer, applied post convolution, aims to minimize the number of parameters and computational demands in the network. A popular pooling method is maximum pooling. Features, once down-sampled by convolution and pooling layers, are relayed to the fully connected layer. Here, the feature maps from the final convolution or pooling layer are flattened into a one-dimensional sequence or vector. These flattened features are connected to one or more fully connected layers, where each neuron holds a learnable weight. The final fully connected layer typically corresponds to the number of classes for classification. Various classifiers can populate this last layer.

In this study, several networks pre-trained on the ImageNet dataset (VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201) serve as deep feature extractors. Spectrogram images of the audio signals are fed into these networks. Deep features are the attributes extracted from CNN layers preceding the classification layer.

The VGG16 architecture was conceived by Simonyan and Zisserman for the ILSVRC 2014 contest. It contains thirteen convolutional and three fully connected layers, summing up to forty-one layers including Maxpool, fully connected layer, ReLu layer, Dropout layer, and Softmax layer. The model expects input images of 224 × 224 × 3 pixels, with its 'fc7' layer (comprising 4094 neurons) designated for deep feature extraction.

ResNet, with its several variations, triumphed in the 2015 ILSVRC contest with a minimal error rate. For this study, the ResNet18, ResNet50, and ResNet101 variants were employed, possessing 18, 50, and 101 layers respectively. Their 'fc1000' layer, holding 1000 fully connected layers, is used for deep feature extraction.

SqueezeNet, a compact alternative to AlexNet, boasts roughly 50 times fewer parameters and a 3x speed improvement. Comprising an independent convolution layer, eight fire modules, and a final convolution layer, its input layer accommodates images of size 227 × 227 × 3. The 'pool10' layer of SqueezeNet, containing 1000 neurons, serves as the feature extractor.

DenseNet, a more recent architecture introduced in 2017, bears similarities to ResNet with nuanced differences. The DenseNet201 model used in this research has direct connections to all successive layers, with the 'conv5_block16' layer designated for feature extraction. Image input size standards were set to 224 × 224 × 3 for networks like VGG16, ResNets, and DenseNet201, and 27 × 227 × 3 for SqueezeNet.

A visual representation of the proposed feature extraction process is depicted in Figure 4, while scatter plots of the deep features sourced from ResNet101 can be found in Figures 5-7.

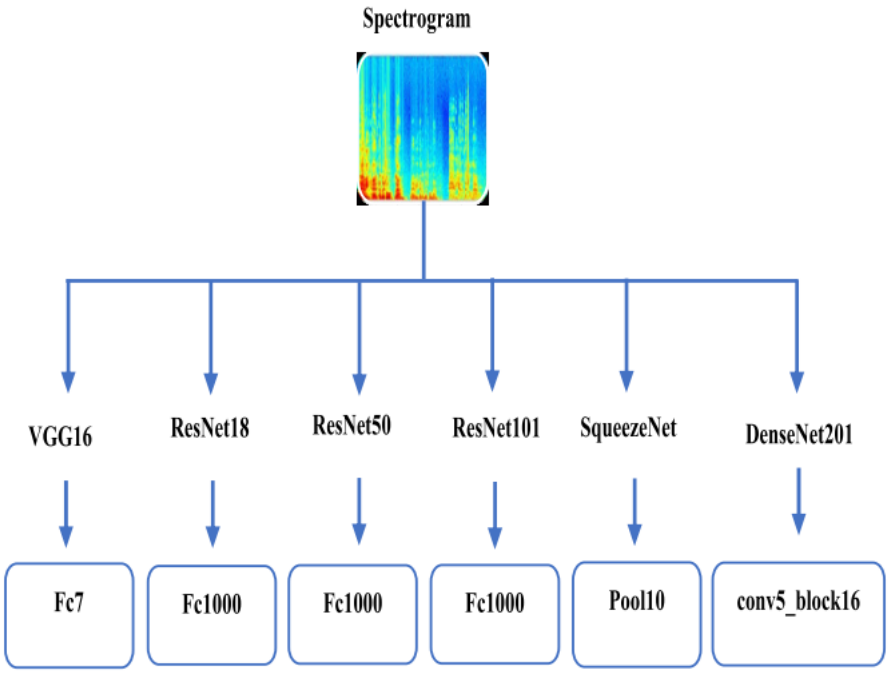


FIGURE 4. Deep feature extraction in pre-trained networks.

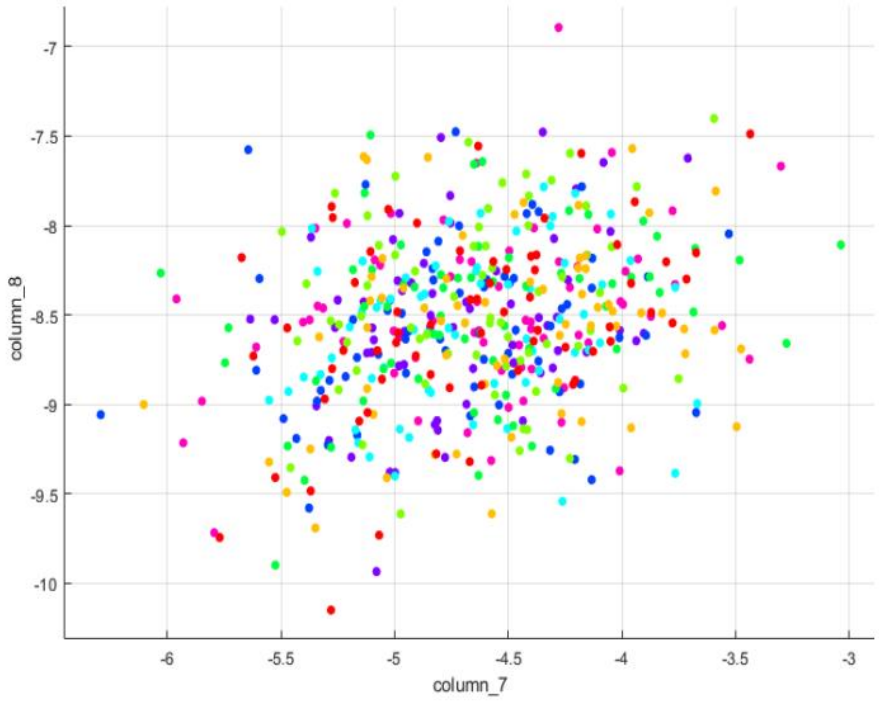


FIGURE 5. 2D representation of deep features for RAVDESS.

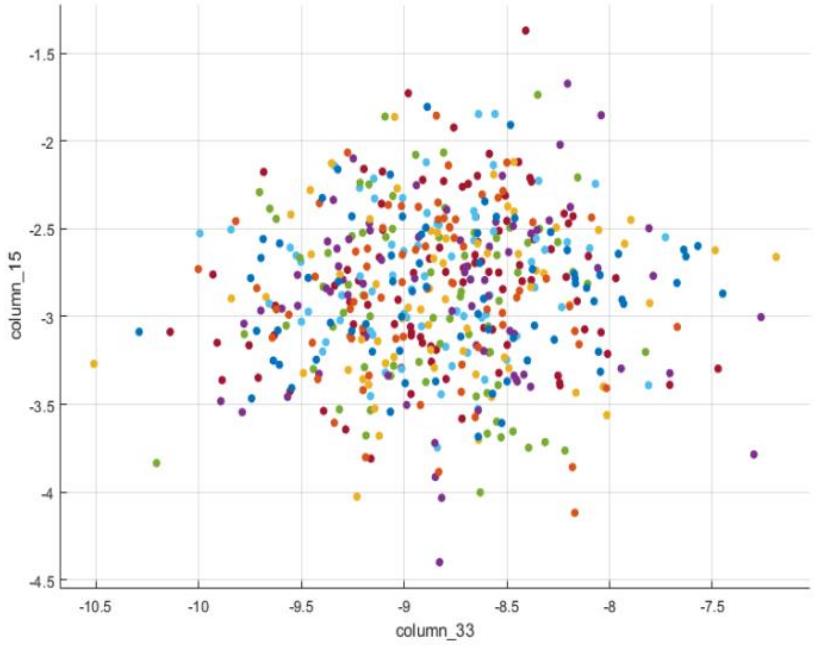


FIGURE 6. 2D representation of deep features for EMO-DB.

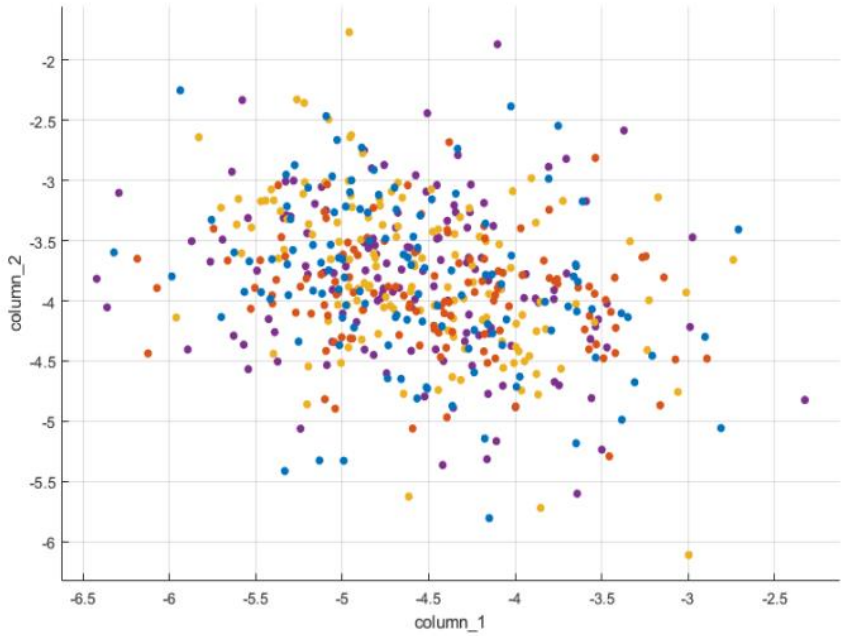


FIGURE 7. 2D representation of deep features for IEMOCAP.

F. FEATURE SELECTION WITH RELIEFF

The ReliefF feature selection algorithm was employed in this research to cherry-pick impactful hybrid features. Feature selection is pivotal in machine learning and data mining domains. ReliefF, lauded for its efficacy in multiple feature selection scenarios, stands as an enhancement of the foundational Relief statistical model. The Relief approach involves sampling from a dataset and then conducting feature selection by devising a model rooted in the proximity of the chosen sample to others in its class and its separation from samples in different classes. The algorithm's weight coefficient update formula is then defined.

$$W[K] = W[K_0] - \frac{\sum_{J=1}^{K} diff\ (A, x_i, H)}{mk} + \sum_{C \neq Class(x_j)} \frac{p\,(C)}{1 - p\,(Class\,(x_i))} \cdot \frac{\sum_{j=1}^{k} diff\ \left(A, x_i, M_j\,(C)\right)}{mk} \tag{8}$$

The pseudocode detailing the ReliefF-based feature selection process is presented in Algorithm 2. Features extracted from various layers of the network models - VGG16's fc7, ResNet's fc1000, SqueezeNet's pool10, and DenseNet201's conv5_block16 - were amalgamated with acoustic features. This composite set then underwent the ReliefF feature selection algorithm. A comparative list of the number of features both before and after the application of the feature selection, specifically for the RADVESS dataset, is tabulated in Table 1.

Algorithm 2 Feature Selection With ReliefF Input:

1. f: Feature vector for train examples

Output:

1. w: Predicted features

Algorithm:
1. set all weights w[K]: = 0;
2. for i := 1 to do
3. randomly select an sample
4. find k-nearest hits sj
5. for each class C class(zi) do
6. for K: = 1 to k do
7. W [K] = W [K0] − PK J=1 diff (A,xi,H) mk +
8. P C6=Class(xj) p(C) 1−p(Class(xi)). Pk j=1 diff(A,xi,Mj(C)) mk
9. end for
10. end for

TABLE 1. Number of features before and after feature selection.

Features	Number of Features Before Feature Selection	Number of Features After Feature Selection
VGG16 {fc7}+Acoustic features	4128	3248
ResNet18 {fc1000}+Acoustic features	1032	756
ResNet50 {fc1000}+Acoustic features	1032	843
ResNet101 {fc1000}+Acoustic features	1032	642
SqueezeNet {pool10}+Acoustic features	1032	794
DenseNet201{conv5_block16}+Acoustic features	1440	965

F. CLASSIFICATION

The culmination of the emotion recognition process is the classification phase. This involves training the classifier with the combined acoustic and deep features extracted. While the Softmax function is typically associated with the classification layer of pre-trained models, this study opted to utilize the Support Vector Machine (SVM) for classification. SVM, introduced by Vapnik in 1995, is adept at tackling both linear and nonlinear problems, often yielding superior results for a plethora of practical applications. Its primary objective is to establish a separating hyperplane amidst data from two distinct classes.

This hyperplane, which can range from two-dimensional to multi-dimensional, creates parallel zones by fashioning two concurrent lines, partitioning the space linearly and in a straightforward manner. SVM's strategy

centers around finding the most expansive margin between categories, aiming to separate them with as broad a gap as feasible. A wider gap between the two classes in SVM typically correlates with enhanced classification performance.

Given an input vector {xi, i = 1, · · · , n}, each should belong to a class, either yi{−1, 1}. The defining equation for a hyperplane in this context would be: (Note: The exact equation wasn't provided in your input).

$$w_0 x + b_0 = 0 \qquad (9)$$

In this context, ww represents the weight vector, xx stands for the input vector, and bb is the bias term. For a particular pair of ww and bb, data can be linearly separated if either of the following conditions is met:

$$w.x_i + b \geq 1 \quad if\ y_i = 1 \qquad (10)$$
$$w.x_i + b \leq 1 \quad if\ y_i = -1 \qquad (11)$$

To address a nonlinear problem using a linear classifier, the kernel method is employed. Here, the input data is transformed into a higher-dimensional space using a function, often denoted as ϕϕ. Within this context, the kernel function KK is defined as follows:

$$k\left(x, x'\right)) = \left(\Phi(x), \Phi(x')\right) \qquad (12)$$

-Linear

$$k\left(x_i, x_j\right) = x_i.x_j \qquad (13)$$

Polynomial

$$k\left(x_i, x_j\right) = \left(x_i.x_j + 1\right)^d \qquad (14)$$

Here, d is the degree of the polynomial.

RESEARCH RESULT

A. DATASET

Three voice datasets used in emotion recognition studies were employed in this research: RAVDESS, EMO-DB, and IEMOCAP. RAVDESS consists of audio-visual recordings of 24 actors portraying eight emotions. In EMO-DB, there are 535 German audio outputs, divided into seven emotional categories. IEMOCAP offers both improvised and scripted audio-visual data from ten actors across ten emotional categories. For the sake of consistency, only four emotions (angry, happy, neutral, and sad) from IEMOCAP were considered in this study.

B. EXPERIMENTAL RESULTS

The method was implemented on a machine equipped with an i7 2.50GHz processor, 12GB RAM, and an NVIDIA 940M GPU using MATLAB R2018a. This study extracted deep features from various pre-trained networks like VGG16, ResNet variations, SqueezeNet, and DenseNet201. Subsequently, acoustic features were combined with the deep features to create thirteen distinct hybrid feature vectors.

For classification, a linear kernel SVM was used. Using 10-fold cross-validation, the datasets were split into test and training sets. The performance of the proposed method is documented in tables, with the RAVDESS dataset achieving a top accuracy of 79.41% using ResNet101 combined with acoustic features. For the EMO-DB dataset, the highest accuracy rate reached 90.21% with the same combination. Meanwhile, the IEMOCAP dataset's best result was 85.37% using VGG16 combined with acoustic features. The training computational complexities of these CNNs are detailed in another table.

TABLE 2. Classification results on RAVDESS dataset.

Feature Vector	Accuracy %
Acoustic features	66.42
DenseNet201{conv5_block16}	75.43
DenseNet201{conv5_block16}+Acoustic features	76.97
DenseNet201{conv5_block16}+Acoustic features with ReliefF	77.46
ResNet18 {fc1000}	70.25
ResNet18 {fc1000}+Acoustic features	75.38
ResNet18 {fc1000}+Acoustic features with ReliefF	74.56
ResNet50 {fc1000}	73.24
ResNet50 {fc1000}+Acoustic features	76.86
ResNet50 {fc1000}+Acoustic features with ReliefF	78.26
ResNet101 {fc1000}	73.77
ResNet101 {fc1000}+Acoustic features	77.56
ResNet101 {fc1000}+Acoustic features with ReliefF	79.41
SqueezeNet {pool10}	74.23
SqueezeNet {pool10}+Acoustic features	73.46
SqueezeNet {pool10} + Acoustic features with ReliefF	75.81
VGG16 {fc7}	71.15
VGG16 {fc7}+Acoustic features	73.36
VGG16 {fc7}+Acoustic features with ReliefF	74.41

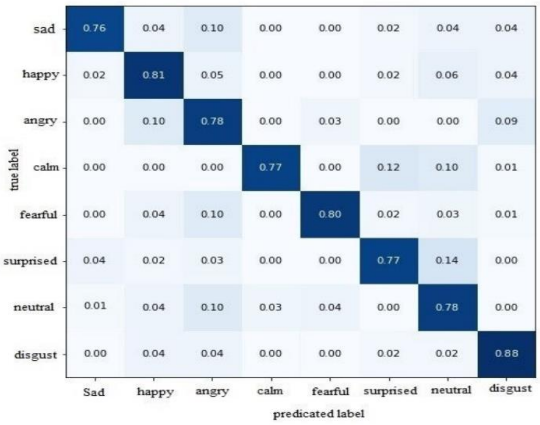


FIGURE 8. Confusion matrix for the RAVDESS dataset.

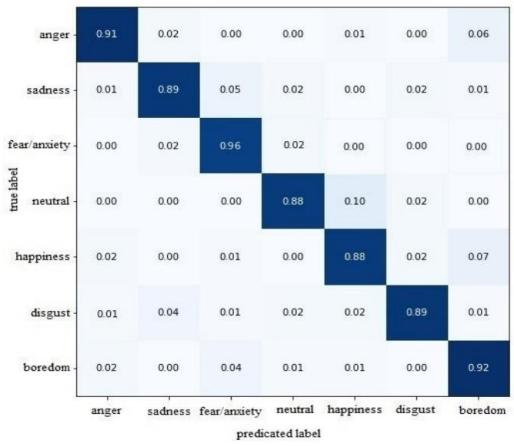


FIGURE 9. Confusion matrix for the EMO-DB dataset.

TABLE 3. Classification results on EMO-DB dataset.

Features	Accuracy %
Acoustic features	76.34
DenseNet201{conv5_block16}	85.42
DenseNet201{conv5_block16}+Acoustic features	87.29
DenseNet201{conv5_block16}+Acoustic features with ReliefF	86.48
ResNet18 {fc1000}	85.06
ResNet18 {fc1000}+Acoustic features	85.47
ResNet18 {fc1000}+Acoustic features with ReliefF	85.42
ResNet50 {fc1000}	83.28
ResNet50 {fc1000}+Acoustic features	87.43
ResNet50 {fc1000}+Acoustic features with ReliefF	88.67
ResNet101 {fc1000}	85.58
ResNet101 {fc1000}+Acoustic features	88.98
ResNet101 {fc1000}+Acoustic features with ReliefF	90.21
SqueezeNet {pool10}	86.24
SqueezeNet {pool10}+Acoustic features	87.46
SqueezeNet {pool10}+Acoustic features with ReliefF	88.63
VGG16 {fc7}	85.27
VGG16 {fc7+Acoustic features	89.21
VGG16 {fc7+Acoustic features with ReliefF	90.12

TABLE 4. Classification results on IEMOCAP dataset.

Features	Accuracy %
Acoustic features	71.36
DenseNet201{conv5_block16}	77.51
DenseNet201{conv5_block16}+Acoustic features	80.46
DenseNet201{conv5_block16}+Acoustic features with ReliefF	82.37
ResNet18 {fc1000}	75.50
ResNet18 {fc1000}+Acoustic features	82.46
ResNet18 {fc1000}+Acoustic features with ReliefF	83.47
ResNet50 {fc1000}	78.65
ResNet50 {fc1000}+Acoustic features	81.56
ResNet50 {fc1000}+Acoustic features with ReliefF	82.74
ResNet101 {fc1000}	77.72
ResNet101 {fc1000}+Acoustic features	81.45
ResNet101 {fc1000}+Acoustic features with ReliefF	82.36
SqueezeNet {pool10}	81.45
SqueezeNet {pool10}+Acoustic features	83.87
SqueezeNet {pool10}+Acoustic features with ReliefF	82.68
VGG16 {fc7}	81.26
VGG16 {fc7}+Acoustic features	84.52
VGG16 {fc7}+Acoustic features with ReliefF	85.37

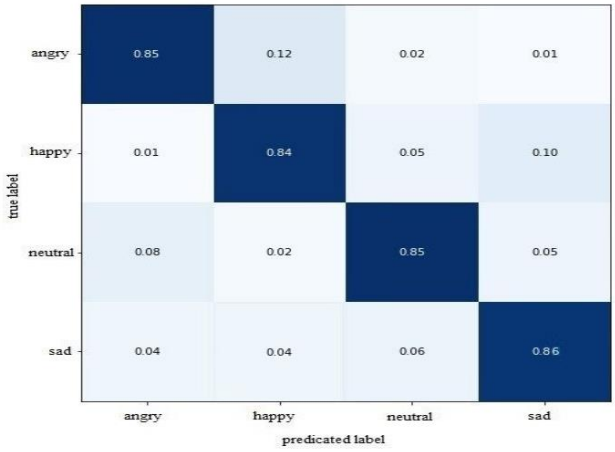


FIGURE 10. Confusion matrix for the IEMOCAP dataset.

TABLE 5. Computational complexity of CNNs.

Models	Training Time(s) for RAVDESS Dataset	Training Time(s) for EMO-DB Dataset	Training Time(s) for IEMOCAP Dataset
DenseNet201	854	364	647
ResNet18	178	96	148
ResNet50	257	101	189
ResNet101	299	134	207
SqueezeNet	237	82	126
VGG16	372	178	246

Experiments were performed using the transfer learning method, solely using spectrogram images. While preserving the original model parameters (excluding the fully connected layers) as initial values, the last layer was adjusted to match the class count in the new dataset. Various hyperparameters were established:

- Minibatch size: 64
- Maximum epoch number: 32
- Learning rate: 1e-4

TABLE 6. Results of transfer learning method.

Models	Accuracy % for the RAVDESS	Accuracy % for the EMO-DB	Accuracy % for the IEMOCAP
DenseNet201	70.86	83.49	81.76
ResNet18	70.86	84.18	80.43
ResNet50	70.46	85.49	82.15
ResNet101	72.34	83.67	82.76
SqueezeNet	71.76	84.73	80.75
VGG16	71.43	85.14	81.45

TABLE 7. ANOVA test results.

Compared Results	P Value for RAVDESS Dataset	P Value for EMO-DB Dataset	P Value for IEMOCAP Dataset
VGG16+Acoustic features & ResNet18+Acoustic features	≈ 0	≈ 0	≈ 0
ResNet18+Acoustic features & ResNet50+Acoustic features	≈ 0	≈ 0	≈ 0
ResNet50+Acoustic features & ResNet101+Acoustic features	0.002	0.002	0.002
ResNet101+Acoustic features & SqueezeNet +Acoustic features	0.141	0.141	0.141
SqueezeNet+Acoustic features & DenseNet201+Acoustic features	0.123	0.118	0.128

According to the results in Table 6:

- Best accuracy for RAVDESS: 72.34% (using ResNet101)
- Best accuracy for EMO-DB: 85.49% (using ResNet50)
- Best accuracy for IEMOCAP: 82.76% (using ResNet101)

The results were statistically analyzed using ANOVA. The findings indicated significant differences between results obtained by different methods, with a "P" value typically near zero.

TABLE 8. Performance comparison with other approaches.

Approach	Features Used	Classifier	Training–Testing Data Splitting	Dataset	Accuracy %
Segokar and Sircar [53]	Continuous wavelet Transform, Prosodic features	SVM	Cross-validation	RAVDESS	60.1
Zeng et al. [54]	Spectrogram	CNN	Cross-validation	RAVDESS	65.97
Bhavan et al. [25]	MFCCs, spectral centroids and MFCC derivatives	Bagged ensemble of SVMs	90% - 10%	RAVDESS	75.69
Proposed model	ResNet101{fc1000}+Acoustic features (Before data augment)	SVM	Cross-validation	RAVDESS	77.26
Proposed model	ResNet101{fc1000}+Acoustic features (After data augment)	SVM	Cross-validation	RAVDESS	79.41
Wang et al. [55]	Fourier parameters, MFCC	SVM (Gaussian kernel)	Not mentioned	EMO-DB	73.3
Kotti et al. [56]	Cepstrum-based features	Linear SVM	Not mentioned	EMO-DB	87.70
Wu et al. [57]	Modulation spectral features (MSFs)	Linear discriminant analysis (LDA)	Cross-validation	EMO-DB	85.84
Proposed model	ResNet101{fc1000}+Acoustic features (Before data augment)	SVM	Cross-validation	EMO-DB	87.68
Proposed model	ResNet101{fc1000}+Acoustic features (After data augment)	SVM	Cross-validation	EMO-DB	90.21
Lee and Tashev [58]	Spectrogram, Frame level features	Recurrent neural network	Cross-validation	IEMOCAP	62.85
Zhao et al. [29]	Mel spectrograms	LSTM	Cross-validation	IEMOCAP	89.16
Proposed model	VGG16{fc7}+Acoustic features (Before data augment)	SVM	Cross-validation	IEMOCAP	82.64
Proposed model	VGG16{fc7}+Acoustic features (After data augment)	SVM	Cross-validation	IEMOCAP	85.37

DISCUSSION

In this study, I explored an emotion recognition system that integrates both acoustic and deep features from speech signals. The use of pre-trained CNN models like VGG16 and ResNet, combined with traditional acoustic features, shows the potential to capture intricate emotion patterns from speech that might be missed if relying solely on either approach. The application of the ReliefF feature selection algorithm further refined the feature set, leading to more efficient and possibly more accurate classification.

However, it's worth noting that while the results were promising, especially when compared to other methods in the literature, there's still room for improvement. The transfer learning approach, although beneficial, might further be optimized with different hyperparameters or even the incorporation of newer models. The use of multiple datasets, including RAVDESS and EMO-DB, also added to the robustness of the research. It's evident that the fusion of deep learning and traditional speech processing techniques offers an intriguing path forward in speech emotion recognition.

CONCLUSIONS AND RECOMMENDATIONS

1. **Integrated Approach**: The study reinforced the effectiveness of integrating both acoustic and deep features. Leveraging pre-trained CNN models with acoustic features can provide a more comprehensive understanding of the emotional nuances in speech signals.

2. **Feature Selection**: The application of the ReliefF feature selection algorithm played a pivotal role. It not only streamlined the feature set but also potentially enhanced the accuracy of the emotion classification.
3. **Comparison with Other Methods**: Results from the study, especially when benchmarked against other methods in the literature, highlighted the efficiency of the proposed method. The accuracy rates achieved, especially on datasets like RAVDESS and EMO-DB, are commendable.
4. **Transfer Learning Potential**: The research illuminated the potential of transfer learning in emotion recognition. Using pre-trained models and fine-tuning them for specific tasks can yield significant benefits, both in terms of computational efficiency and accuracy.

Implementation of Research Results:

1. **Real-world Applications**: The results can be directly applied to real-world scenarios such as customer service to gauge customer sentiment, in mental health applications to monitor patient emotional states, or even in entertainment industries for dynamic content recommendation.
2. **Enhanced Training**: The use of data augmentation techniques, such as adding background noise or time-shifting, can be implemented more broadly to improve the robustness of other voice-based systems.
3. **Optimization**: For practitioners aiming to implement the findings, the research underscores the need for periodic optimization. With the rapid advancements in deep learning architectures, newer models can be tested and integrated to further boost accuracy.
4. **Custom Models**: Based on the success of the transfer learning approach, organizations can invest in creating custom models, trained on industry-specific data, to improve the specificity and relevance of the emotion recognition process.
5. **Tool Integration**: Given the versatility of the models used, businesses can integrate the emotion recognition system into their existing tools, enhancing user experience and gaining valuable insights into customer emotions and sentiments in real-time.

ADVANCED RESEARCH

Limitations:

1. **Dataset Constraints**: The study utilized three datasets - RAVDESS, EMO-DB, and IEMOCAP. While these datasets are comprehensive, they may not encompass all possible emotional nuances or languages, potentially limiting the universality of the results.
2. **Pre-trained Model Limitations**: Relying on pre-trained CNN models means the study was bound by their architectures. These models are designed for general purposes, and while transfer learning is effective, the original training might not be entirely suited for speech emotion recognition.
3. **Complexity**: The combination of acoustic features, deep features, and SVM can introduce computational complexities, potentially making real-time applications more challenging.

4. **Static Emotion Classes**: The classification was based on predefined emotion categories. In real-world scenarios, emotions may not always be so clear-cut, and there could be overlaps.

Suggestions for Further Research:

1. **Diverse Datasets**: Investigating more diverse and larger datasets, including different languages, dialects, and cultural nuances, can provide a more holistic understanding of speech emotion recognition.
2. **Custom Models**: Instead of solely relying on pre-trained models, building custom deep learning architectures tailored for emotion recognition might yield even more accurate results.
3. **Dynamic Emotion Recognition**: Future research could delve into recognizing mixed or transitional emotions rather than static categories. This would mimic real-life scenarios more closely.
4. **Integration with Other Modalities**: Combining audio data with other modalities, like facial expressions or physiological signals, can enhance the accuracy and robustness of emotion recognition systems.
5. **Optimization Techniques**: Implementing optimization techniques, both at the architectural and algorithmic levels, can make the system more suited for real-time applications, especially in devices with computational constraints.

ACKNOWLEDGMENT

I would like to express my profound gratitude and appreciation to all those who provided me with the support and guidance throughout this research journey. I am especially indebted to Mrs. Meenakshi Thalor ma'am and Principal Mr. P. B. Mane sir for their invaluable feedback, mentorship, and encouragement. Their wisdom and insights have been pivotal to the success of this endeavor. Additionally, I'd like to thank my colleagues, friends, and family who have constantly motivated and believed in me. Any achievements and findings presented in this research are a reflection of the collective efforts of all those mentioned.

REFERENCES

A. B. Kandali, A. Routray, and T. K. Basu, ''Emotion recognition from assamese speeches using MFCC features and GMM classifier,'' in Proc. IEEE Region Conf. (TENCON), Nov. 2008, pp. 1–5, doi: 10.1109/tencon. 2008.4766487.

A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, ''Deep features-based speech emotion recognition for smart affective services,'' Multimedia Tools Appl., vol. 78, no. 5, pp. 5571–5589, Mar. 2019, doi: 10.1007/s11042-017-5292-7.

A. Milton, S. Sharmy Roy, and S. Tamil Selvi, ‘‘SVM scheme for speech emotion recognition using MFCC feature,’’ Int. J. Comput. Appl., vol. 69, no. 9, pp. 34–39, May 2013, doi: 10.5120/11872-7667.

D. Połap, ‘‘Model of identity verification support system based on voice and image samples,’’ J. Univers. Comput. Sci., vol. 24, pp. 460–474, Jan. 2018.

D. V. Waghmare, R. Deshmukh, P. Shrishrimal, and G. Janvale, ‘‘Emotion recognition system from artificial Marathi speech using MFCC and LDA techniques,’’ in Proc. 5th Int. Conf. Adv. Commun., Netw., Comput., 2014, pp. 1–10. [16] S. Demircan and H. Kahramanli, ‘‘Feature extraction from speech data for emotion recognition,’’ J. Adv. Comput. Netw., vol. 2, no. 1, pp. 28–30, 2014, doi: 10.7763/jacn.2014.v2.76.

F. Chenchah and Z. Lachiri, ‘‘Acoustic emotion recognition using linear and nonlinear cepstral coefficients,’’ Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 11, pp. 1–4, 2015, doi: 10.14569/ijacsa.2015.061119.

G. Lu, L. Yuan, W. Yang, J. Yan, and H. Li, ‘‘Speech emotion recognition based on long short-term memory and convolutional neural networks,’’ J. Nanjing Univ. Posts Telecommun., vol. 38, no. 5, pp. 63–69, Nov. 2018, doi: 10.14132/j.cnki.1673-5439.2018.05.009.

H. Gunes and M. Piccardi, ‘‘Bi-modal emotion recognition from expressive face and body gestures,’’ J. Netw. Comput. Appl., vol. 30, no. 4, pp. 1334–1345, Nov. 2007, doi: 10.1016/j.jnca.2006.09.007.

H.-S. Bae, H.-J. Lee, and S.-G. Lee, ‘‘Voice recognition based on adaptive MFCC and deep learning,’’ in Proc. IEEE 11th Conf. Ind. Electron. Appl. (ICIEA), Jun. 2016, pp. 1542–1546, doi: 10.1109/iciea.2016.7603830.

K. Han, D. Yu, and I. Tashev, ‘‘Speech emotion recognition using deep neural network and extreme learning machine,’’ in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 1–5.

K. R. Malik, M. Ahmad, S. Khalid, H. Ahmad, F. Al-Turjman, and S. Jabbar, ‘‘Image and command hybrid model for vehicle control using Internet of Vehicles,’’ Trans. Emerg. Telecommun. Technol., vol. 31, no. 5, p. e3774, 2019, doi: 10.1002/ett.3774.

M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, ‘‘Primitivesbased evaluation and estimation of emotions in speech,’’ Speech Commun., vol. 49, nos. 10–11, pp. 787–800, Oct. 2007, doi: 10.1016/j.specom.2007.01.010.

M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, ‘‘Recognizing emotions induced by affective sounds through heart rate variability,’’ IEEE Trans. Affect. Comput., vol. 6, no. 4, pp. 385–394, Oct. 2015, doi: 10.1109/TAFFC.2015.2432810.

P. Schlegel, S. Kniesburges, S. Dürr, A. Schützenberger, and M. Döllinger, ‘‘Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings,’’ Sci. Rep., vol. 10, no. 1, p. 10517, Jun. 2020, doi: 10.1038/s41598-020- 66405-y.

S. Mittal, S. Agarwal, and M. J. Nigam, ‘‘Real time multiple face recognition: A deep learning approach,’’ in Proc. Int. Conf. Digit. Med. Image Process. (DMIP), 2018, pp. 70–76, doi: 10.1145/3299852.3299853.

T. Anvarjon, Mustaqeem, and S. Kwon, ‘‘Deep-net: A lightweight CNNbased speech emotion recognition system using deep frequency features,’’ Sensors, vol. 20, no. 18, p. 5212, Sep. 2020, doi: 10.3390/s20185212.

V. Garg, H. Kumar, and R. Sinha, ‘‘Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers,’’ in Proc. Nat. Conf. Commun. (NCC), Feb. 2013, pp. 1–5, doi: 10.1109/ncc.2013.6487987.

Y. Huang, G. Zhang, X. Li, and F. Da, ‘‘Small sample size speech emotion recognition based on global features and weak metric learning,’’ Acta Acust., vol. 37, pp. 330–338, May 2012. [19] X. Li and M. Akagi, ‘‘Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model,’’ Speech Commun., vol. 110, pp. 1–12, Jul. 2019, doi: 10.1016/j.specom.2019.04.004.