SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I4P103

Volume 12 Issue 4, 19-35, April 2025

Original Article

A Novel Approach of Speech Stress Emotion Recognition Using Visualized Image of Metrices

Vaijanath V. Yerigeri[1], Seema V. Yerigeri[2]

[1]*Department of Electronics & Computer Engineering, M. B. E. S. College of Engineering, Ambajogai, Maharashtra, India.*
[2]*Department of Mechanical Engineering, M. B. E. S. College of Engineering, Ambajogai, Maharashtra, India.*

[1]*Corresponding Author : vaijanatha.y@gmail.com*

Received: 02 February 2025 Revised: 04 March 2025 Accepted: 05 April 2025 Published: 29 April 2025

Abstract *- Algopsychalia is inimical to the host. The present lifestyle of homo sapiens is stressful, due to which they suffer from psychogenic pain. Psychologists warn humans about algopsychalia's destructive form, i.e. stress. Excessive stress can trigger suicidal tendencies in a person. Stress and emotions are highly co-related; therefore, the paper proposes efficient detection of stress-related emotions using speech to identify the level of stress and intimacy prior to the threat of suicidal ideations. The paper explores cepstral coefficient-based perceptual features like Mel Frequency, Inverted Mel frequency, Gammatone Wavelet, Gammatone Frequency, Perceptual Linear Predictive, Bark Frequency and Revised Perceptual Linear Prediction. Features are represented as an image and are input to the learning model. Representing features as an image and applying a Region-based Convolutional Neural Network (R-CNN) learning algorithm for evaluating auditory cues is the novelty of the proposed work. R-CNN learning reduces computational costs. The performance of a system is analyzed with the help of a benchmark dataset specific to stress, i.e., SUSAS. Comparative analysis is presented to demonstrate improvement in Speech Emotion Detection (SED) performance. The overall accuracy of 90.66% of stress-related emotions is achieved.*

Keywords *- Speech Emotion Recognition (SER), Gammatone Wavelet Cepstral Coefficients (GWCC), Revised Perceptual Linear Prediction (RPLP), Bark Frequency Cepstral Coefficients (BFCC), Perceptual Linear Predictive coefficients (PLPC), Gammatone Frequency Cepstral Coefficients (GFCC).*

1. Introduction

'The human brain uses speech for enhancing and organizing cognition in the form of interior monologue' - was a motivating force for many researchers to work in speech perception and production. According to Kramer, speech affects not only conveys emotional and physical state of the speaker but also personality, intelligence and appearance [1]. It also reflects the age and gender of the speaker [2]. Machine-man Interface (MMI) or Human Computer Interface (HCI) success rate depends upon the accuracy with which it can detect speech and emotions [3, 4]. Speech Emotion Recognition (SER) is a complex task [5].

Negative or positive emotions either change Heart Rate (HR), sub-glottal pressure and Blood Pressure (BP) and affect the depth of respiratory movements, resulting in Speech Impairment (SI). Thus, emotions control the Sympathetic Nervous System (SNS) and Para Sympathetic Nervous System (PSNS) [17].

Facial expressions and signals like Electroencephalogram (EEG) are alternatives for Emotion Recognition (ER), but speech proved to be a strong candidate as it is non-invasive, natural, low cost and remotely analyzed. Advanced medical diagnosis uses SER to detect Alzheimer's, Parkinson's, Voice Stress and mental disorders [6, 7]. Researchers are working on detecting psychiatric disorders using speech [8-14]. People of every age group face stress, maybe due to workload or lifestyle. Stress directly affects mental and physical health (diabetes, asthma, depression, fatigue, acid peptic disease and laryngeal tremors) [15]. Medical science defines stress as a silent killer. Unbearable stress may result in impulsive action, viz., committing suicide. Therefore, detecting stress to save lives is the desideratum. In the present scenario, a person is away from home, which rules out facial expression analysis for ER and favors SER.

Understanding emotions and extracting related features are two major tasks. Schlosberg presentedthe first 3D model, which consists of activation, valence, and potency (energy) in emotion space [16]. These three parameters represent a person's strong disposition, positive/negative emotions and energy in speech, respectively. Carl introduced the concept of *Affective Computing* [17]. They explored temporal features (pauses, articulation rate, speech rate), fundamental frequency and average speech spectrum. The discrete categorical model proposed by Ekaman covered 8 emotions, viz. surprise, anger, happiness (joy), neutral, sad, disgust, fear (anxiety) and boredom [18]. Murray and Rainer demonstrated a correlation between emotions and utterance timing, utterance pitch contour, and voice quality [19, 20]. Cowie proposed a 2D model (activation, valence) for empirical analysis [21, 22].

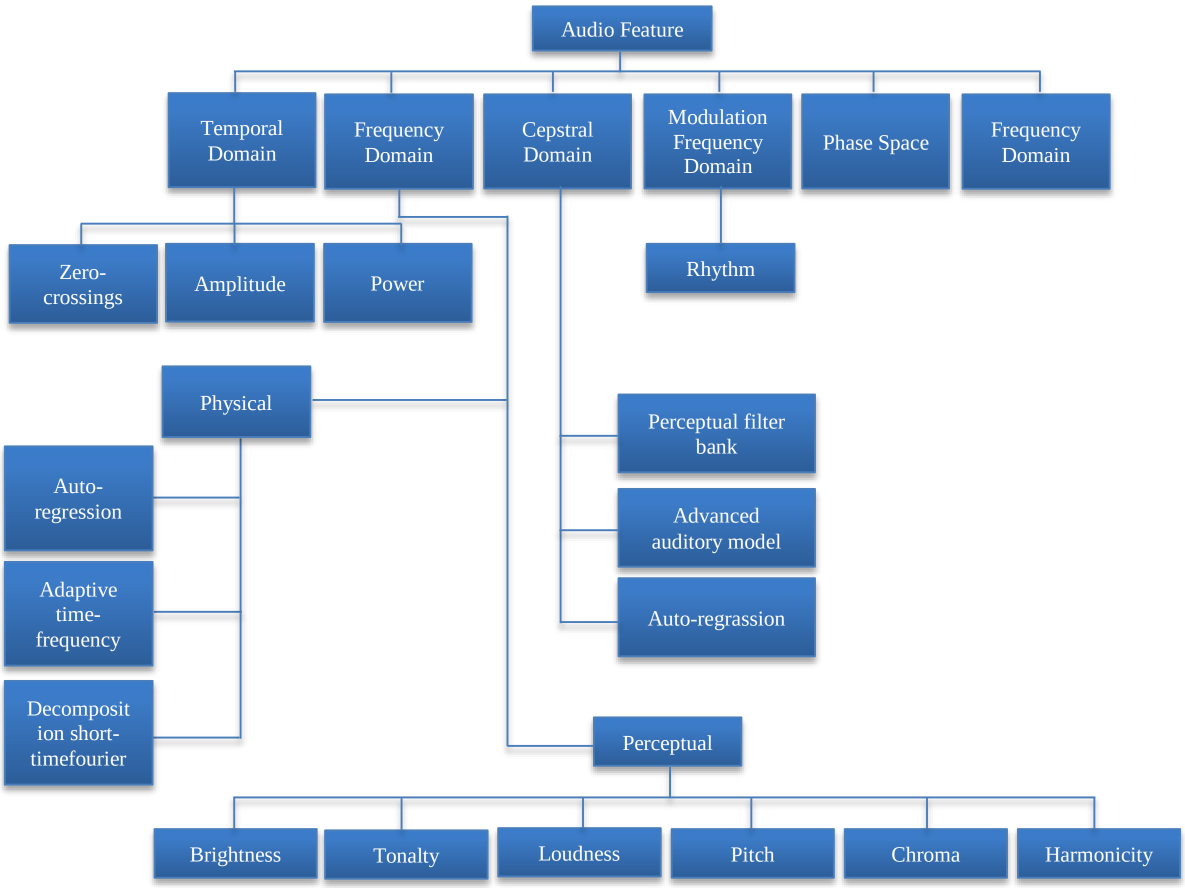


Fig. 1 Audio feature set [102]

In Figure 1 features are presented as different 'Domain' (temporal, frequency, cepstral, modulation, phase and Eigenvalues). Temporal, frequency and cepstral-based domains have enticed researchers to develop SER.

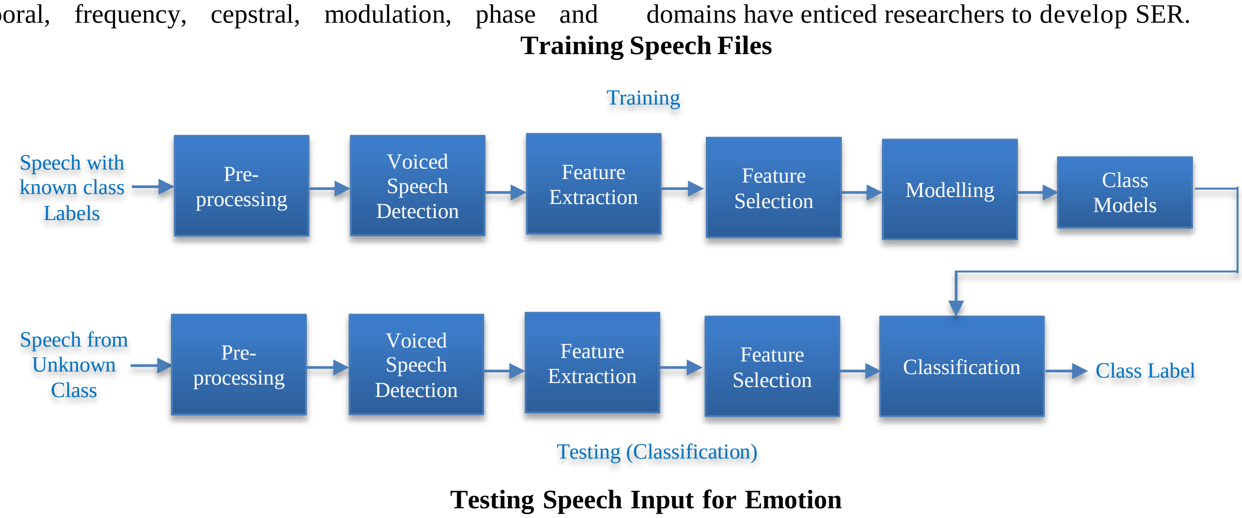


Fig. 2 SER block diagram

In Figure 2, the SER model is based on a machine learning algorithm, which requires training of features extracted from the speech signal. Speech (audio) signal is segmented into overlapped/non- overlapped frames using the window function. Features shown in Figure 1 are extracted and presented in vector form. The feature vector is input to Machine Learning Algorithms (MLA) for training purposes. Training generates Trained Model (TM) viz., *.NET*. In the testing phase, audio inputs other than those used for training are used. Input may be from a dataset or real-time signal whose features are extracted and given to the classifier. The classifier model accepts feature vectors of new audio input, compares them with TM and produces a class of emotions. MLA performance relies on feature vectors and the classifier used. Researchers have used different classifier techniques, viz. Kernel Regression (KR) and Machine Learning Control (MLC), Double Sparse Learning (DSL) [23], Hidden Markov Model - HMM [24], Support Vector Machine (SVM) [25], Artificial Neural Network (ANN) [26], K-means Neural Network (KNN) [27], Deep Belief Networks (DBN) [28].

Robust and accurate stress analysis being a major goal; in this paper, the author proposes novelty at two levels, i.e. feature vector generation and classifier. Homo sapiens' loudness and frequency response is non-linear and alters as per the age, health and surrounding environment. Therefore, Perceptual Features (PF) are selected to generate feature vectors based on Cepstral Coefficients (CC). CC is based on non-linear frequency warping, thus the best candidate for the SER system. At the classifier level, a novel method of analysing speech emotion visually to classify verities of emotion families is being introduced.

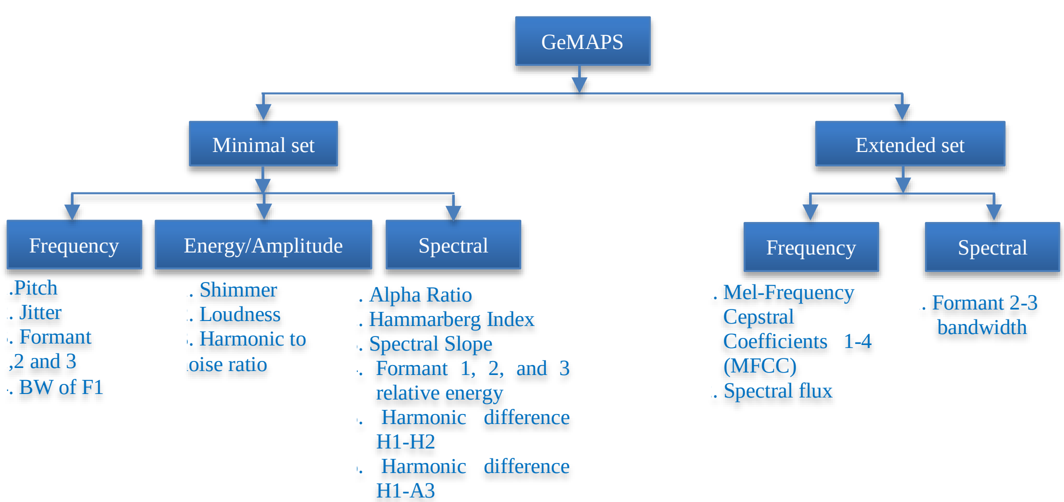
The proposed model constructs the SER process as an image classification problem. It converts the extracted PF into an image called Image Matrices. Representing metrices as a visualized image explores all the state-of-the-art techniques developed for classifying an image. Constructed images are not complex to the level of RGB, instead limited to gray scale images. Region-based Convolutional Neural Network (R-CNN) classifier is used in MLA.

The main contribution of the research work is representing Perceptual feature metrices as an image. The same has been trained using the R-CNN classifier. To the best of our knowledge, the above combination, i.e. image representation of perceptual features and CNN classifier, has not been evaluated by researchers.

A benchmark dataset, SUSAS (Speech Under Simulated and Actual Stress), is used to evaluate system performance, and the author created a database in Marathi (Indo-Aryan) language. Reasons for selecting the Marathi language are given as follows.

US outsources, on average 65% of Information Technology (IT) business to India, out of which Maharashtra acquires the major IT business. Maharashtrian people speak Marathi. Thus, there is a high probability that IT professionals in Maharashtra, due to heavy workload, may be affected by stress. This motivated the author to explore the SER system for the Marathi language.

The next part of the paper is organized as follows. Section 2 describes work done by researchers in this area. The summary and challenges related to SER are described in Section 3. In this section, a problem statement is formulated. The proposed work model is presented in Section 4. Feature vector formation using perceptual features and the mathematical model of the CNN classifier are described in this section. Experimental setup and related results are presented in Sections 5 and 6, respectively. The system performance of the proposed model is presented in Section 7.



NOTE: Minimal set with extended set total 88 parameters

Fig. 3 GeMAPS parameters

2. Related Work

The section presents the work done related to SER. The year 2015 was at a crossroads for feature extraction techniques [29] introduced toolbox for speech i.e. Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and shared software library 'openSMILE' – a baseline work may be used by researchers. Minimal feature set and extended feature set are two options available. Figure 3 depicts both the sets. INTERSPEECH 2009 Emotion Challenge was based on the use of the OpenSMILE toolkit.

C.K. Yogesh generated a feature set using OSBSBC (OpenSmile Bi-Spectral and Bi-Coherence). Wrapper-based Particle Swarm Organization (PSO) Biogeography-Based Optimization (BBO) algorithm and Multi-Cluster Feature Selection (MCFS) techniques were used to extract significant features, respectively, for 2016 and 2017. Optimization techniques like Hybrid BBO and BBO with PSO are used to reduce feature vector sets. SER performance analysis was done using BES, SUSAS and Surrey Audio-Visual Expressed Emotion (SAVEE) dataset, using SVM and Extreme Learning Machine (ELM) classifier [30, 31].

Zhang generated a feature set using eGeMAP and ComParE. They classified 9 emotions using Multi-Task Deep Networks (MTDN) [32]. Wang proposed Deep Neural Network (DNN) trained using utterance-level features for classification of gender, emotions and age [33]. Siddique Latif extracted features of the dataset, which had five corpus and three languages, using eGeMAP [28]. They derived that training small chunks of data but with a greater number of languages will enhance SER performance, and Deep Belief Network (DBN) outperforms the SVM and Sparse Auto Encoder (SAE) classifier.

Few researchers didn't use OpenSMILE and proposed a SER system with different feature sets and classifiers. Yuan Zong proposed feature set optimization and selecting significant speech segments using Double Sparse Learning (DSL) [23]. They followed a pyramid structure with 384 features as a base and then optimized further. With the SVM classifier, they achieved 62.6% and 33.42% for the eNTERFACE and Acted Facial Expressions in the Wild (AFEW) datasets, respectively.

The real Life Depression and Affect Recognition (ALDAR) theme was presented by the Audio Visual Emotion Challenge (AVEC 2017). Fabien proposed the use of the toolbox Collaborative Voice Analysis Repository (COVAREP) (v1.3.2) for feature extraction and identifying sentiments of human beings [35].

The next section summarizes the literature review in table format and discusses challenges with an outline of the proposed work.

3. Summary and Challenges-Formulating Problem Statement

Figure 4 represents a literature survey summary. This section discusses the challenges related to the work carried out. OpenSMILE toolbox was introduced in 2015, but after that, it has not been updated as per the introduction of new feature extraction techniques in different domains. Surrey reveals that the usage of standard auditory toolboxes may end up with huge feature vectors that may have redundant information. In most cases, the huge database may 'confuse' the training network.

Therefore, researchers focused on optimizing feature vectors without compromising system performance, i.e. accuracy. At this juncture, two approaches popped up. The first approach uses optimization algorithms to reduce feature vectors and identify significant features. In the second approach, the researcher identifies the prominent features and manually selects a number of features. Verities of features were used but without identifying decisive perceptual features.

The user proposes representing the feature vector as an image and using the classifier techniques that are proven and specific to image analysis. The next section describes challenges in SER and the reason for representing the feature set as an image.

3.1. Motivation: Representing Metrics as an Image

Figure 5 depicts variations in emotion. A few important points are explored, listed as follows,

1. Emotions themselves are complex. On top of that, correlating emotions with specific speech features is still complex.
2. Emotions are dynamic or obscure in nature.
3. There is no explicit division of features with respect to emotion.
4. Acoustic- related features are highly affected by parameters viz., content of speech, speaking rate, speaking style and varieties of speakers.
5. Multiple emotions are correlated with the same utterance. Segmenting speech utterances precisely is a complex task.
6. Emotions are highly influenced by the speaker's culture, environment, accent, etc.
7. Emotional state may be transient or long-term.

The listed complexity demands a technique that can accommodate the dynamic nature of emotion. A novel method is required which can visualize feature vectors and the variations in the same.

The proposed work tries to address points – 4, 5 and 6 from the mentioned challenges. At the same time, we should also track similarities in the images. A detailed explanation of the same is covered in Section5.

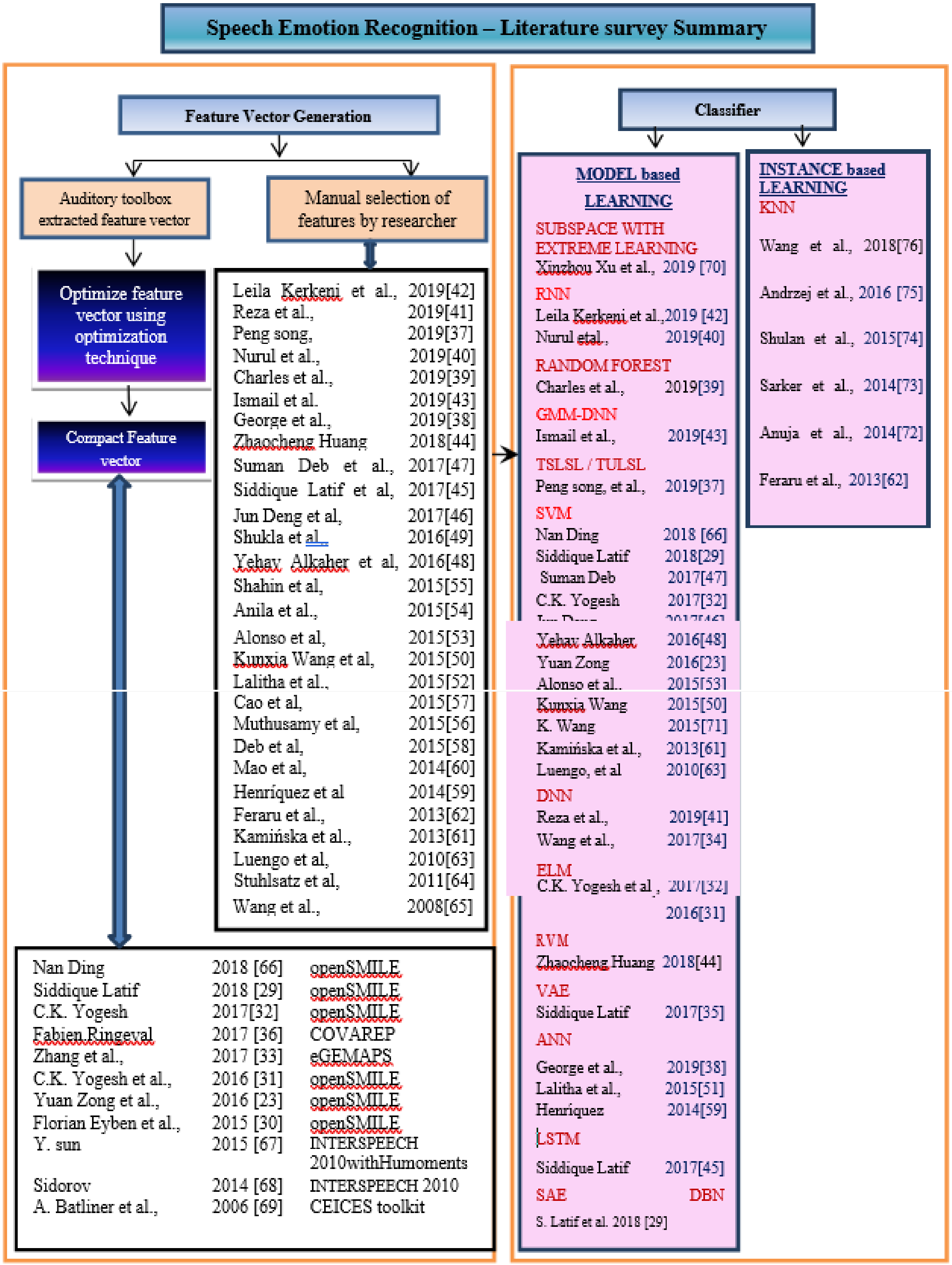


Fig. 4 Schematic presentation of literature review- work done by verities of researchers

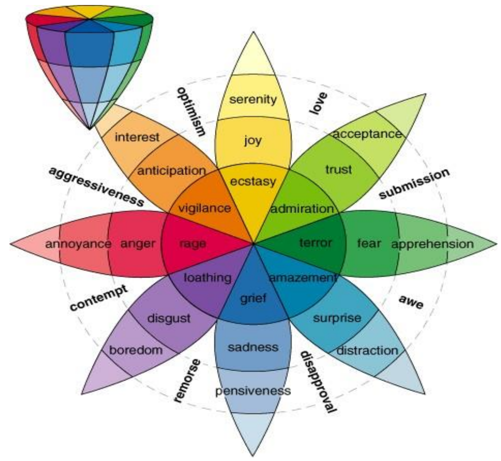


Fig. 5 Emotion variation

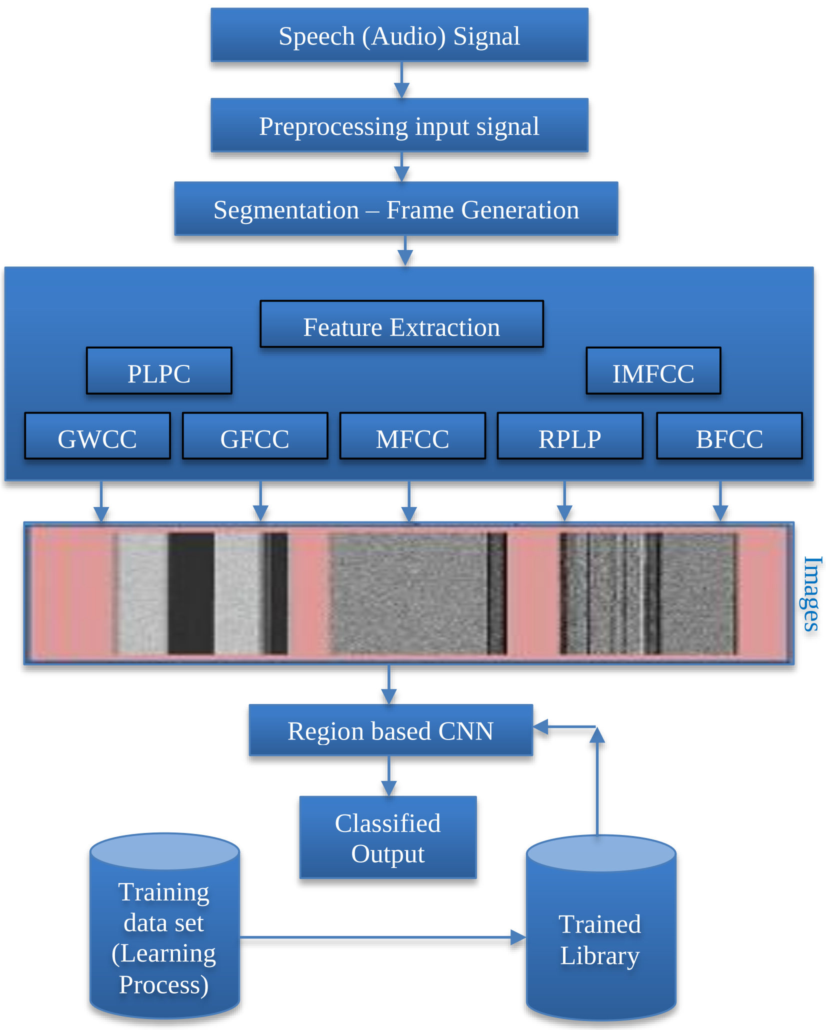


Fig. 6 Architecture diagram: proposed model

Referring to Figure 4, there are two types of classifiers, viz. Eager learners and Lazy, i.e. instance-based learners. Eager learners are mostly preferred as training time is more, but testing time is less. However, the author proposes a classifier that operates on the *'Region'* of an image, and the same is not utilized for the SER system. The next section discusses the proposed model.

4. SER: Proposed Model

Referring to Figure 6, there are three important modules of SER, viz. speech signal pre-processing, generating feature vector, and classifier, which are discussed as follows.

4.1. Pre-Processing Module

In a real-time environment, environmental or surrounding noise will be present. The author created a database, and the SUSAS corpus is recorded in a noisy environment. The first step involves normalization and removal of DC offset. After that, background noise removal is done using a Wiener filter. Fairly clean speech is then segmented into several frames using the Hamming windowing technique w(n). Boundary ripples and discontinuities are catered by this technique. The window is 30ms with approximately 30% overlap i.e. 10ms.

$$y(n) = S(n) * w(n) \tag{1}$$

Where n is ranging from 0 to (N-1), S(n) = input signal

$$w(n) = 0.54 - 0.46\cos\left(\frac{2*\pi*n}{N-1}\right), 0 \le n \le (N-1) \tag{2}$$

4.2. Feature Extraction

This section describes the verities of features extracted to generate a set. The proposed work generates a feature vector set of perceptual features with energy, pitch and fundamental frequency. Details are as follows,

4.2.1. Analyzing Gammatone Wavelet Cepstral Coefficient (GWCC)

Gammatone function models auditory response. Patterson proved the same experimentally. Cochlea is modeled by a parallel bank of band pass filter [76]. Yang and Slaney demonstrated that the cochlea and Gammatone Cepstral Coefficient transfer function, which maintains constant Q is the same [77, 78]. It was observed by Solbach and Mallat that affine Wavelet Transform (WT) exactly model cochlea as it has constant Q behavior [79, 80]. Thus, the primary observation is that Gammatone Function (GF) with WT will improve SER system performance. Thus, theoretical studies indicate positive about SER performance enhancement if WT is combined with Gammatone Function (GF).

$$\text{GDF(t)} \quad = \quad f(t) = t^{N-1} e^{-\beta t} e^{j\omega c^{t}} u(t) \tag{3}$$

Where, $GDF(t)$ = $f(t)$= $gamma\ distribution\ function$, β = $bandwidth\ parameter$, $u(t)$ = $step\ function$, N = $order\ of\ rise\ or\ decay\ function$ $(3 \le N \le 5)$ and jw$_c$ = $centre\ frequency\ (rad|sec)$

Gamma Distribution Function (GDF) is modulated on sinusoids,

Fourier Transform (FT) of $GD(t)\ is$

$$\hat{G}DF(\omega) = f(t) is\ \hat{f}(\omega) = \frac{(N-1)!}{(\beta + j(\omega - \omega_c))^N} \tag{4}$$

ω = $Angular\ frequency\ (rad|sec)$,
∴ Gammatone Wavelet Function (GWF),

$$\hat{\Psi}(\omega) = f(\omega) * \hat{f}(\omega) = \frac{j\omega * (N-1)!}{(\beta + j(\omega - \omega_c))^N} \tag{5}$$

$\hat{\Psi}(0) = 0$, therefore, time domain function will be,

$$\psi(t) = \frac{d}{dt}\left\{t^{N-1} e^{-\beta t} e^{j\omega_c t} u(t)\right\} \tag{6}$$

$$\Psi(t) = \left\{(N-1)t^{N-2} + \alpha * t^{N-1}\right\} e^{\alpha t} u(t) \tag{7}$$

Where $\alpha = \beta + j\omega_c$

[81] proposed that in Equation 4 if the derivative order of the numerator is less than the denominator order, then GF higher order derivatives can be approximated to Wavelet.

Gammatone Wavelet Filter Bank (GWFB)

Patterson [82] proved that the human auditory system filter characteristics perfectly match the order four of the GF impulse response. Giasberg defined auditory filter function, i.e. Equivalent Rectangular Bandwidth (ERB), as follows,

$$ERBfunction(f) = 24.7(1 + 4.37 * 10^{-3} f) = 24.7 + \frac{f}{9.26} \tag{8}$$

Bandwidth (BW) of fourth-order Gammatone function = 1.019 ERB.

Centre frequency fc for channel k is calculated as,

$$f_c(k) = (f_{\max} + 228.83) * \left(e^{k * \frac{\log\left(\frac{c + f_{\min}}{c + f_{\max}}\right)}{K}} \right) - 228.83 \tag{9}$$

Where,
$f_c(k)$ = $Center\ frequency$
K = $Number\ of\ filters\ in\ filter\ bank$ $0 < k < K$
f_{max} = $Higher\ cutoff\ frequency$, $Typically\ 4KHz$
f_{min} = $Lower\ cutoff\ frequency$, $Typically\ 20Hz\ to\ 133Hz$

GF ERB scale has practically logarithmic characteristics. According to [83], the $fc(k)$ function is equallydistributed on the scale.

4.2.2. MFCC Filter Bank

The ear has nonlinear filter characteristics i.e. High Frequency (HF) has a smaller number of filter stages as compared to the Low Frequency (LF) range. MFCC analysis is based on human perception of speech. Mel space filter (triangular filter) bank is the base for RPLP and MFCC. The MFCC algorithm in Table 1 is presented as follows,

Table 1. MFCC algorithm

	$S(n)$ = Time domain speech signal, $S_i(n)$ = Segmented speech (i number of frames)
1	Compute complex DFT (C-DFT) $S_i(K)$ for each and every frame
2	Calculate the Power spectrum, $P_i(K) = \frac{1}{N}\lvert S_i(K)\rvert^2$
3	Apply triangular filter bank to $P_i(K)$.
4	Filter bank energy is computed by multiplying all. Filter Bank (FB) with the spectrum, and finally, coefficients are summed up.
5	Linear Frequency scale is mapped to Mel Frequency scale, $mel(f) = 1125 * \ln\left(1 + \frac{f}{700}\right)$
6	Calculate log of output $mel(f)$
7	Compute Discrete Cosine Transform (DCT) of log filter bank energies

4.2.3. BFCC Filter Bank and PLPC

To calculate BFCC, the DCT of the logarithm of the spectrum of Bark Frequency Filter Bank is computed. The equation for warping Bark Scale Frequency is as follows,

$$bark(f) = 13 * \tan^{-1}\left(\frac{0.76 * f}{1000}\right) + 3.5\tan^{-1}\left(\frac{f}{7500}\right) \quad (10)$$

To calculate PLPC, the IDFT of the bark scale is computed, and then the linear prediction is applied for spectralsmoothness.

4.2.4. Flow Chart to Compute Perceptual Features

Figure 7 presents a flow chart for calculating perceptual features.

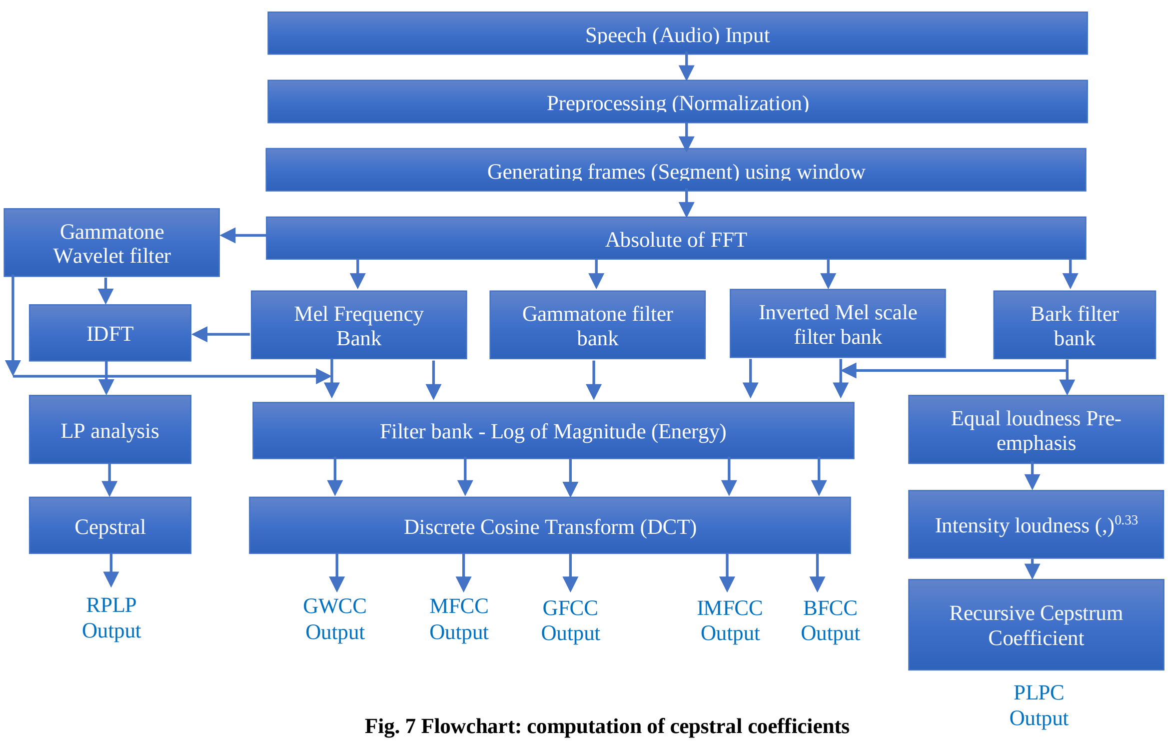


Fig. 7 Flowchart: computation of cepstral coefficients

4.2.5. Additional Features

This group covers features such as pitch, energy, and vocal tract frequency.

1. Pitch Detection

Table 2. Pitch detection

1	Calculate STFT of S(n) (frequency domain), $F_t(f)=\sum_{\alpha=1}^{N} S_{\alpha,t}\delta(f-\alpha f_0)+N_t(f)$ *$N_t(f)$ = Noise spectrum Power density, $f_t(f)$= STFT, f_0 represents frequency* t = time, N= Harmonic, $S_{\alpha,t}$ = power of Nth harmonic
2	Calculate log spaced PSD of every frame. $F_t(\varsigma)=\sum_{\alpha=1}^{N}\varsigma(n)\delta(\varsigma-\log(\alpha)-\log(f_0))+N_t(\varsigma)$ $where, \varsigma=\log f$
3	Impulse response, $H(\varsigma)=\sum_{\alpha-1}^{N}\delta(\varsigma-\log(N))$
3	Convolution operation is performed between $F_t(\zeta)$ and $H(\zeta)$ $F_t(\zeta)*H(\zeta).\prod$
4	Noise suppression is achieved by using broaden peak filter stage which has impulse response, $h_{bp}(\zeta)=r-\log(u-\cos(2*\prod * e^{\zeta}))$ $with\ limit$ $\log(0.5)<\zeta<\log(0.5+N)$
5	Noise component(s) with high energy narrow band may dominate output of a filter. Therefore,spectrum of each frame is compressed. $F'_t(\zeta) = F_t(\zeta)^{\beta_t(\zeta)}$, $t = time\ index$, $\beta_t(\zeta) = compression\ index$
6	To formulate $\beta_t(\zeta)$ smoothed spectrum ($\bar{F}_t(\zeta)$) is calculated taking into consideration two cases, viz., without noise and with noise.
7	Compute $\bar{F}_t(\zeta)$, $F_t(\zeta)$ is passed through Low Pass Filter (LPF) stage in two domains viz. time and log frequency domain.
8	Function $\bar{F}_t(\zeta)$, may be approximated to long term average spectrum Byrne [87] and Brookes [88]. $\therefore\ \bar{F}_t(\zeta),\approx NN_t(\zeta)$.
9	Point 8 approximation results in Compression Index, $\beta_t(\varsigma)=\frac{\log NN_t(\varsigma)}{\log F_t(\varsigma)}$, $\overline{F_t}(\varsigma) normalisedto NN_t(\varsigma)$.
9	Convolve $F'_t(\zeta)$ and $h_b(\zeta)$.
10	*Pitch of voice is nothing but the resulting highest peak (maximum peak) in the feasible range.*

Pitch detection is an appropriate technique to model a single quasi-periodic signal, viz., speech and music [84, 85]. Pitch can be detected using time, frequency and both domains. Gonzalez proposed a Pitch Estimation Filter Robust to High Levels of Noise (PEFAC) algorithm for robust pitch detection [86]. The algorithm involves Power Spectrum Density (PSD) calculation of harmonic in log scale and use of broaden peaks filter and spectrum compression to eliminate noise consisting of high amplitude narrowband component. The same is summarized in Table 2.

2. Energy

Speech signal energy is computed using the Teager Energy Operator (TEO).

$$\Psi\left[S(n)\right]=S^2(n)-S(n-1)*S(n+1) \tag{11}$$

3. Vocal Tract Frequency (VTF)

VT is a closed cylinder with a 17 cm to 18 cm length. It generates a set of formant frequencies that change as the articulator introduces vowel sounds. These frequencies are different for men and women.It also changes with emotions.

4.3. Classification Module-Region based CNN Classifier

Refer to Figure 8. Visualized images created by vectors of verities of feature coefficients represent different regions.

As the sizes of the coefficients selected are fixed, the region size, as well as the position in an image,of the feature will be fixed. Thus, the R-CNN network, which is efficient in region-based detection, is a good candidate for the SER system.

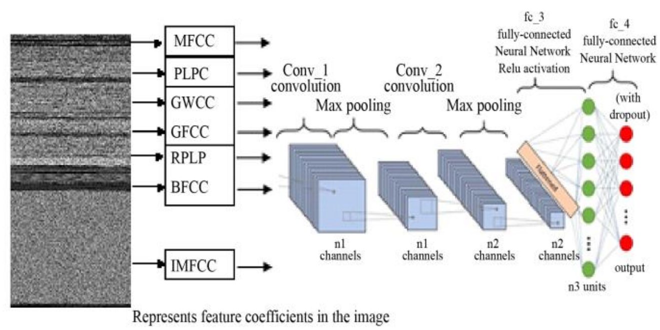
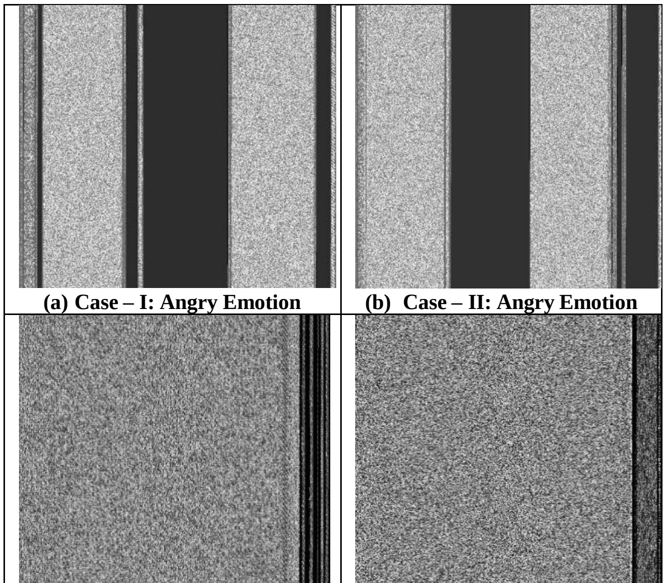


Fig. 8 Architecture diagram: proposed model

Figures 9(a), (b) and (c), (d) represent a single class of emotion (Angry and Sad, respectively). For the same emotion, the pattern is almost the same but with observable changes in the same.

Angry and sad emotional images are distinct. Thus, R-CNN may be trained (Table 3) for such similarities in an image. The next part elaborates on R-CNN.



(a) Case – I: Angry Emotion (b) Case – II: Angry Emotion (c) Case – III: SAD Emotion (d) Case – IV: SAD Emotion

Fig. 9 (a), (b)Angry, and (c), (d) Sad.

4.3.1. R-CNN Algorithm

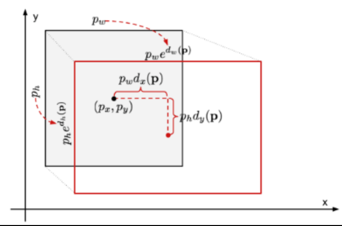
R-CNN is a pioneering approach that applies deep models to region/object detection [89]. It first selects several proposed regions from an image. Anchor box may be one of the selection techniques. Regions are then labeled for categories and bounding boxes (Region of Interest (RoI)).

Features of proposed regions are extracted by forward computation, performed using CNN (Table 4). The next part elaborates on the R- CNN process.

Table 3. Train CNN Network to classify images

1	Select N number of ROI of different sizes per image using selective search. The ROI may be category-independent and of the target object.
2	Apply warping to region candidates to generate a fixed-size template.
3	Total ROI will be $N + 1$ as background is considered no object of interest.
4	Fine-tune CNN having $N + 1$ classes. Adjust smaller learning rates.
5	Each one forward propagation generates a feature vector for every ROI.
6	Train each class independently and create binary SVM.
7	Set overlap threshold (Intersection over Union – IoU)
8	Binary SVM uses feature vector generated in step-6 and overlap threshold to decide positive or negative sample.
9	The regression model is trained using CNN features to set right the predicted detection window on bounding box correction offset

Table 4. Bounding box regression

1	$Coordiante\ for\ predicted\ bounding\ box\ \boldsymbol{p} = input\ to\ transformation\ functions = (p_x, p_y, p_w, p_h),$ $where\ p_x, p_y\ are\ centre\ coordianate, p_w\ is\ width\ a$ $p_h\ is\ height,$
2	$Coordiante\ for\ Ground\ truth\ box\ \boldsymbol{g} = (g_x, g, g_w, g),$
	$if\ (IoU\ between\ p\ and\ g \gg 0.6\ )\ then$
3	Configure regressor for learning scale-invariant the transformation between two centres and log scale the transformation between heights and width.
4	Calculate transformation. $\hat{g}_x = p_w d_x(\boldsymbol{p}) + p_x$, $\hat{g}_y = p_h d_y(\boldsymbol{p}) + p_y$, $\hat{g}_w = p_w\, e^{d_w(\boldsymbol{p})}$, $\hat{g}_h = p_h\, e^{d_h(\boldsymbol{p})}$ 
5	$Bounding\ box\ Correction\ factor\ \ d_i(\boldsymbol{p}).\quad i \in \{x, y, w, h\},$ $and\ can\ take\ any\ value\ ranging\ from\ \{-\infty, \infty\}$
6	Learning Targets are: $t_x = \frac{(g_x - p_x)}{p_w}$, $t_y = \frac{(g_y - p_y)}{p_h}$, $t_w = \log\left(\frac{g_w}{p_w}\right)$, $t_h = \log\left(\frac{g_h}{p_h}\right)$
7	Minimizing the SSE loss with regularization: $l_{reg} = \sum_{i \in \{x,y,w,h\}} (t_i - d_i(\boldsymbol{p}))^2 + \lambda * \lVert \boldsymbol{W} \rVert^2$
8	Select best λ using cross-validation.
9	end

5. Experimental Setup

Robust Speech Processing Laboratory has a publicly distributed SUSAS database specifically foranalyzing stress-based speech. The database has four partitions. The audio signal is sampled at 8 KHz using 16-bit ADC. The database contains 16,000 utterances recorded by 32 speakers of the age group 22 to 76 years. It contains long, real-time speech recorded by Apache helicopter pilots. Benchmark Marathi speech database is not commonly available [90]. Researchers created their own data set [91, 92] but haven't distributed the same publicly. The author created their own database with the help of nonprofessional speakers, as SER is supposed to be used by common people. Rode NT 2A microphone and Behringer x32 digital mixer were used. Parameter settings are described in Table 5.

Table 5. Parameters considered for database creation

Sr.	Parameter	Specifications
1	Stage of life (age bar / number of speakers)	Old (60+ / 1), middle (31 to 55 / 1), young (20 – 30 / 5), teenage (13 – 15/

		5) and child (7 – 12 / 2)
2	Emotion states	5 emotions, viz. sad, happy, neutral, surprised and angry
3	Gender	M / F
4	Sentence / Repetition rate	size 5 - 7 word(s) / 3
6	Variations in statements	10
7	Recording Duration	2 / 3 sec
8	Bit rate	16 / 192 kbps
9	Number of bits	16
10	Sampling rate	44.1 KHz
11	Channel	Mono
12	File format	mp3

Feature vector formation is based on Table 6.

Table 6. Coefficients for feature vector formation

Sr.	Parameter	Coefficients	Sr.	Parameter	Coefficients
1	GWCC	35	2	GFCC	25
3	BFCC	30	4	RPLP	10
5	MFCC	15	6	IMFCC	15
7	PLPC	11	8	Energy	1
9	Pitch	1	10	Vocal tract Fundamental frequency	1

The feature vector size is 144, which forms the matrix of 12 x 12 matrix. For training there will be 30 speech files of each emotion. Therefore,

$$total\ matrix\ for\ training = number\ of\ speech\ files * number\ of\ emotions * feature\ vector.$$

Training the network is done with an input matrix of size 12 x 12 x 1 (H x W x Dimension). The first convolutional layer has Kernel (K) selected as a 3 x 3 x 1 matrix, stride length = 1 and the same padding is configured in the software. Thus, the input matrix is augmented to 13 x 13 x 1, and the kernel applied to this produces an output matrix of 10 x 10 x 1, as per the formula ((image size – kernel size)/stride +1) = ((12 – 3)/1+1).

Max pooling is used instead of average pooling to get better output. Max pooling uses a 3 x 3 x 1 matrix with stride length = 1. The output matrix size is ((10 – 3)/1 + 1) = 8. We kept the same parameters for convolutional layer 2. So, the convolutional output matrix size was 6 x 6, and the max pool layer matrix size was 4 x 4.

The output from the max pool from the layer is input to a Fully Connected Network (FCN) having 4096 neurons. Two FCN networks were used.

6. Experimental Results and Discussion

The section explores variations in features with different emotions. Emotions considered are viz. Sad, angry, surprised, happy and neutral. Figure 10 represents a ribbon plot of features viz. MFCC, BFCC, GFCC, GWCC, PLPC, IMFCC, RPLP, energy, pitch and vocal tract frequency for Marathi speech database.

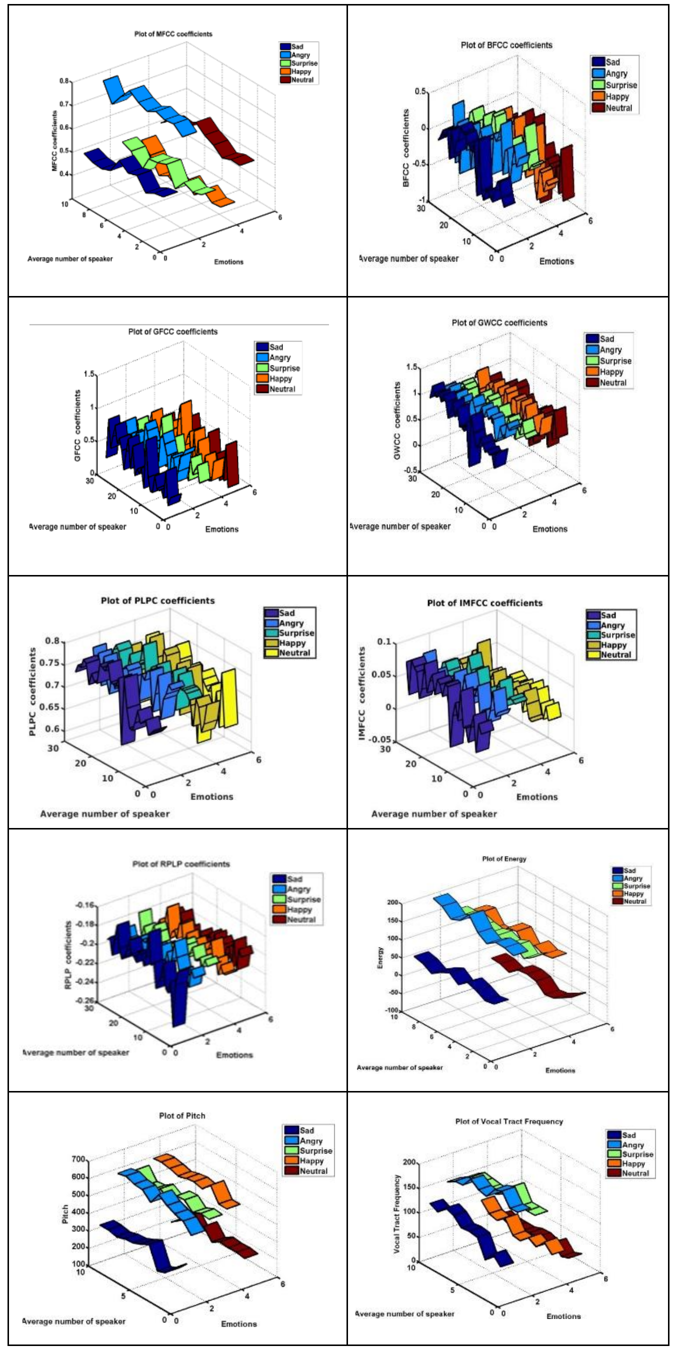


Fig. 10 Feature's ribbon plot

The feature vector set is divided into perceptual features (70%) and non-perceptual features (30%). Figure 11 depicts the overall feature set distribution.

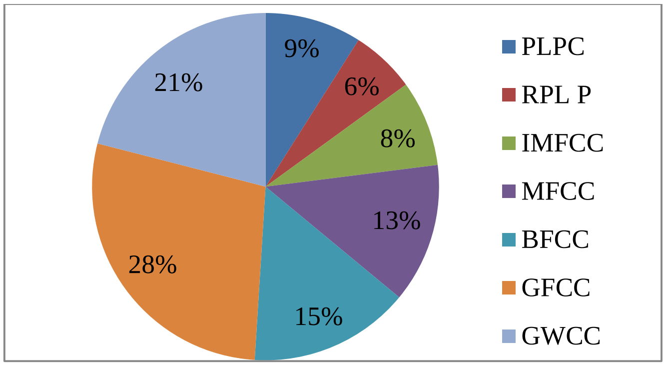


Fig. 11 Pie chart: feature vector distribution

7. Performance Evaluation

SER system accuracy is calculated as,

$$Accuracy = \frac{Number\ of\ correct\ emotions\ detected}{Total\ number\ of\ samples\ of\ that\ emotion}\ X\ 100$$

Confusion matrix is depicted in Tables 7 to 8 of the proposed scheme for Marathi speech and the SUSAS database, respectively. Sad and angry emotions are related to stress. The average accuracy for the same is 93.5%. The overall accuracy of SER for the Marathi database is 90.6%.

Table 7. Results with marathi database

Emotions	Neutral	Happy	Surprise	Angry	Sad
Neutral	89	4	1	2	4
Happy	6	85	8	1	0
Surprise	0	3	92	5	0
Angry	0	1	4	95	0
Sad	5	0	0	3	92

SUSAS database has Slow, Soft, Angry, Question and fast speech for stress tests. Refer to Table 8. Stress is closely related to angry and fast speech and sometimes may be in questionable mode. These threespeaking styles are detected with an average accuracy of 91.66%.

Table 8. Results with SUSAS database

Emotions	Fast	Question	Slow	Angry	Soft
Fast	91	3	0	6	0
Question	5	88	1	5	1
Slow	0	3	80	0	27
Angry	1	3	0	96	0
Soft	0	4	9	0	87

As the Region-based CNN classifier is not used by other researchers, the author compared the results (Table 9) with the learning algorithm presented by researchers, as follows.

Table 9. Comparative performance of the proposed system with existing systems

Author	Year	Feature set	Classifier	Database	Accuracy
Wei Jiang et al.	2019 [93]	eGemaps, MFCCs, IS10	SVM Improved Shared-Hidden -Layer Auto encoder (SHLA)	IEMOCAP (audio-Visual dat)	65%
Peng et al.,	2019 [94]	1,582 features from the openSMILE toolbox, including 34 acoustic low-level descriptors (LLDs)	Transfer supervised linear subspace learning and transfer unsupervised linear subspace learning	Berlin, eInterface(audio)	74.5%
YeSim Ülgen Sonmez	2019 [95]	MFCC, LPC, PLP	Subspace Discriminant Analysis	EMODB (audio)	87.1%
Leila et al.,	2019 [41]	MFCC and Modulation Spectral (MS)	RNN	Berlin Spanish	92%
Jianyou Wang et al.,	2019 [96]		Dual-Sequence LSTM	IEMOCAP	73.3%
Noushin et al.,	2019 [97]	MFCC, pitch, intensity	3D-CNN	SAVEE (audio) RML (audio) eINTERFACE' 05	81.05% 77.0% 72.33%
Badshah et al.,	2017 [98]	Spectrograms	Convolutional Neural Network (CNN)	Berlin	84.3%
Feraru et al.,	2017 [61]	MFCC, PARCOR, LAR	KNN	SROL (audio)	75.83%
Wootaek et al.,	2016 [99]	STFT	CNN & RNN	Berlin	86.86%
Jun Deng et al.,	2016 [45]	Phase based features	SVM	Geneva Whispered Emotion Corpus (audio)	72.5%
Zhengwei	2014 [100]	affect-salient features	Semi-CNN	SAVEE	63.3%

et al.,				EMO-DB (audio)DES (audio) MES (audio)	81.4% 74.6% 76.1%
Qironget al.,	2014 [59]	affect-salient features	CNN	SAVEE EMO -DBDESMES	60,7% 78.3% 75.8% 69.9%
Li et al.,	2013 [101]	MFCC	DNN and HMM	NIST RT03S Fisher (audio)	77.92%
Proposed Scheme		Perceptual Features	Region-Based CNN	SUSAS (audio) Marathi Speech Database (audio)	91.66% 90.6%

The IEMOCAP database has a combination of audio and video. The audio and video feature sets are huge, and SER accuracy decreases. Only audio depicts better performance, which is more appropriate in distant communication.

8. Conclusion

The paper presents the novel idea of representing feature vectors as an image, which opens a wide range of techniques specific to image analysis. Visual inspection of feature-constructed images exploresthe influence of emotions on the feature set. Marathi database (author-created) and SUSUS (benchmark database) were used to analyze system performance. The combination of perceptual features with R-CNN itself is a unique technique presented by the author for SER, and the same has resulted in better performance, i.e. analyzing stress-related emotions like surprise, anger and sad. As shown in Table 5, neutral and happy emotions were also considered for efficient classification. MFCC, GFCC, GWCC and BFCC are more prominent discriminators. Table 6 indicates that the 'Anger' emotion (SUSAS database) is detected with 96% accuracy. Stressful people normally speak very fast with anger and often have questionable states as the situation is not in control from their perspective. These three emotions are detected with an average accuracy of 91.6%. Further research work may be extended to use the Faster R-CNN, Masked R-CNN, and You Only Look Once (YOLO) classifiers. The SER system will be more useful for common people when a mobile app is created and real-time emotion detection is displayed while the telephonic conversation is going on.

References

[1] Ernest Kramer, "Judgment of Personal Characteristics and Emotions from Nonverbal Properties of Speech," *Psychological Bulletin*, vol. 60, no. 4, pp. 408-420, 1963. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ö. Özgür Bozkurt, and Z. Cihan Tayşı, "Audio-Based Gender and Age Identification," *22nd Signal Processing and Communications Applications Conference*, Trabzon, Turkey, pp. 1371-1374, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[3] Rafael A. Calvo, and Sidney D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18-37, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[4] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[5] Jianhua Ma et al., "Ubiquitous Intelligence and Computing," *Third International Conference Proceedings, Lecture Notes in Computer Science*, Wuhan, China, pp. 1-1190, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[6] Theodoros Kostoulas et al., "Affective Speech Interface in Serious Games for Supporting Therapy of Mental Disorders," *Expert Systems with Applications*, vol. 39, no. 12, pp. 11072-11079, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[7] Karmele López-de-Ipiña et al., "On the Selection of Non-Invasive Methods based on Speech Analysis Oriented to Automatic Alz-heimer Disease Diagnosis," *Sensors*, vol. 13, no. 5, pp. 6730-6745, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[8] Nicola Vanello et al., "Speech Analysis for Mood State Characterization in Bipolar Patients," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Diego, CA, USA, pp. 2104-2107, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[9] Gillinder Bedi et al., "A Window into the Intoxicated Mind? Speech as an Index of Psychoactive Drug Effects," *Neuropsychopharmacology*, vol. 39, pp. 2340-2348, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] Gillinder Bedi et al., "Automated Analysis of Free Speech Predicts Psychosis Onset in High-Risk Youths," *NPJ Schizophrenia*, vol. 1, pp. 1-7, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[11] Zahi N. Karam et al., "Ecologically Valid Long-Term Mood Monitoring of Individuals with Bipolar Disorder Using Speech," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, pp. 4858-4862, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[12] Amir Muaremi et al., “Assessing Bipolar Episodes using Speech Cues Derived from Phone Calls,” *Pervasive Computing Paradigms for Mental Health*, pp. 103-114, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[13] Agnes Grünerbl et al., “Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140-148, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[14] Venet Osmani, “Smartphones in Mental Health: Detecting Depressive and Manic Episodes,” *IEEE Pervasive Computing*, vol. 14, no. 3, pp. 10-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[15] James Z. Zhang et al., “Analysis of Stress in Speech Using Adaptive Empirical Mode Decomposition,” *Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, pp. 361-365, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[16] Harold Schlosberg, “Three Dimensions of Emotions,” *Psychological Review*, vol. 61, no. 2, pp. 81-88, 1954. [CrossRef] [Google Scholar] [Publisher Link]

[17] Carl E. Williams, and Kenneth N. Stevens, “Emotions and Speech: Some Acoustic Correlates,” *The Journal of the Acoustical Society of America*, vol. 52, no. 4, pp. 1238-1250, 1972. [CrossRef] [Google Scholar] [Publisher Link]

[18] Paul Ekman, “An Argument for Basic Emotions,” *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169-200, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[19] Iain R. Murray, and John L. Arnott, “Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion,” *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097-1108, 1993. [CrossRef] [Google Scholar] [Publisher Link]

[20] Rainer Banse, and Klaus R. Scherer, “Acoustic Profiles in Vocal Emotion Expression,” *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614-636, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[21] R. Cowie et al., “Emotion Recognition in Human-Computer Interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[22] Roddy Cowie, and Randolph R. Cornelius, “Describing the Emotional States that are Expressed in Speech,” *Speech Communication*, vol. 40, no. 1-2, pp. 5-32, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[23] Yuan Zong et al., “Double Sparse Learning Model for Speech Emotion Recognition,” *Electronics Letters*, vol. 52, no. 16, pp. 1410-1412, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[24] J.F. Gómez-Lopera et al., “The Evaluation Problem in Discrete Semi-Hidden Markov Models,” *Mathematics and Computers in Simulation*, vol. 137, pp. 350-365, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[25] Li Zhiyong et al., “Reject Inference in Credit Scoring Using Semi-Supervised Support Vector Machines,” *Expert Systems with Applications*, vol. 74, pp. 105-114, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[26] Jase D. Sitton, Yasha Zeinali, and Brett A. Story, “Rapid Soil Classification Using Artificial Neural Networks for Use in Constructing Compressed Earth Blocks,” *Construction and Building Materials*, vol. 138, pp. 214-221, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[27] Jesus Maillo et al., “kNN-IS - An Iterative Spark- Based Design of the K-Nearest Neighbors Classifier for Big Data,” *Knowledge-Based Systems*, vol. 117, pp. 3-15, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[28] Siddique Latif, Rajiv Rana, and Junaid Qadir, “Adversarial Machine Learning and Speech Emotion Recognition: Utilizing Generative Adversarial Networks for Robustness,” a*rXiv*, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[29] Florian Eyben et al., “The Geneva Minimalistic Acoustic ParameterSet (GeMAPS) for Voice Research and Affective Computing,” *Transactions on Affective Computing*, vol. 7, no. 2, pp. 190-202, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[30] C.K. Yogesh et al., “A New Hybrid PSO Assisted Biogeography-Based Optimization for Emotion and Stress Recognition from Speech Signal,” *Expert Systems with Applications*, vol. 69, pp. 149-158, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[31] C.K. Yogesh et al., “Hybrid BBO_PSO and Higher Order Spectral Features for Emotion and Stress Recognition from Natural Speech,” *Applied Soft Computing*, vol. 56, pp. 217-232, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[32] Yue Zhang et al., “Multi-Task Deep Neural Network with Shared Hidden Layers: Breaking Down the Wall between Emotion Representations,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 4990-4994, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[33] Zhong-Qiu Wang, and Ivan Tashev, “Learning Utterance-Level Representations for Speech Emotion and Age/Gender Recognition Using Deep Neural Networks,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, pp. 5150-5154, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[34] Siddique Latif et al., “Transfer Learning for Improving Speech Emotion Classification Accuracy,” a*rXiv*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[35] Arianna Mencattini et al., “Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models,” *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 314-327, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[36] Peng Song, "Transfer Linear Subspace Learning for Cross-corpus Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265-275, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[37] George M. Slavich, Sara Taylor, and Rosalind W. Picard, "Stress Measurement Using Speech: Recent Advancements, Validation Issues, and Ethical and Privacy Considerations," *The International Journal on the Biology of Stress*, vol. 22, no. 4, pp. 408-413, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[38] Charles R. Marmar et al., "Speech-Based Markers for Posttraumatic Stress Disorder in US Veterans," *Depression and Anxiety*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[39] Nurul Lubis et al., "Positive Emotion Elicitation in Chat-Based Dialogue Systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 866-877, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[40] Reza Lotfian, and Carlos Busso, "Curriculum Learning for Speech Emotion Recognition from Crowdsourced Labels," *IEEE/ACM Transactions on Audio, Speech, And Language Processing*, vol. 27, no. 4, pp. 815-826, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[41] Leila Kerkeni et al., *Automatic Speech Emotion Recognition Using Machine Learning*, Intech Open, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[42] Ismail Shahin, Ali Bou Nassif, and Shibani Hamsa, "Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network," *IEEE Access*, vol. 7, pp. 26777-26787, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[43] Zhaocheng Huang, and Julien Epps, "An Investigation of Partition-Based and Phonetically Aware Acoustic Features for Continuous Emotion Prediction from Speech," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 653-668, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[44] Siddique Latif et al., "Mobile Health in the Developing World: Review of Literature and Lessons from a Case Study," *IEEE Access*, vol. 5, pp. 11540-1156, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[45] Jun Deng et al., "Exploitation of Phase-Based Features for Whispered Speech Emotion Recognition," *IEEE Access*, vol. 4, pp. 4299-4309, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[46] Suman Deb, and Samarendra Dandapat, " Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 360-373, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[47] Yehav Alkaher, Osher Dahan, and Yair Moshe, "Detection of Distress in Speech," *IEEE International Conference on the Science of Electrical Engineering*, Eilat, Israel, pp. 1-5, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[48] Mohinish Shukla, Katherine S. White, and Richard N. Aslin, "Prosody Guides the Rapid Mapping of Auditory Word Forms Onto Visual Objects in 6-Month-Old Infants," *The Proceedings of the National Academy of Sciences*, vol. 108, no. 15, pp. 6038-6043, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[49] Kunxia Wang et al., "Speech Emotion RecognitionUsing Fourier Parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69-75, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[50] S. Lalitha et al., "Time-Frequency and Phase Derived Features for Emotion Classification," *Annual IEEE India Conference*, New Delhi, India, pp. 1-5, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[51] S. Lalitha et al., "Emotion Detection Using MFCC and Cepstrum Features," *Procedia Computer Science*, vol. 70, pp. 29-35, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[52] Jesús B. Alonso et al., "New Approach in Quantification of Emotional Intensity from the Speech Signal: Emotional Temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554-9564, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[53] Anila M. D'Mello, Peter E. Turkeltaub, and Catherine J. Stoodley, "Cerebellar tDCS Modulates Neural Circuits during Semantic Prediction: A Combined tDCS-fMRI Study," *Journal of Neuroscience*, vol. 37, no. 6, pp. 1604-1613, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[54] Ismail Shahin, and Mohammed Nasser Ba-Hutair, "Talking Condition Recognition in Stressful and Emotional Talking Environments Based on CSPHMM2s," *International Journal of Speech Technology*, vol. 18, pp. 77-90, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[55] Hariharan Muthusamy, Kemal Polat, and Sazali Yaacob, "Improved Emotion Recognition Using Gaussian Mixture Model and Extreme Learning Machine in Speech and Glottal Signals," *Mathematical Problems in Engineering*, vol. 2015, no. 1, pp. 1-13, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[56] Houwei Cao, Ragini Verma, and Ani Nenkova, "Speaker-Sensitive Emotion Recognition via Ranking: Studies on Acted and Spontaneous Speech," *Computer Speech & Language*, vol. 29, pp. 186-202, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[57] Suman Deb, and Samarendra Dandapat, "A Novel Breathiness Feature for Analysis and Classification of Speech under Stress," *Twenty First National Conference on Communications*, Mumbai, India, pp. 1-5, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[58] Patricia Henríquez et al., "Nonlinear Dynamics Characterization of Emotional Speech," *Neurocomputing*, vol. 132, pp. 126-135, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[59] Qirong Mao et al., "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[60] Dorota Kamińska, Tomasz Sapiński, and Adam Pelikant, "Comparison of Perceptual Features Efficiency for Automatic Identification of Emotional States from Speech," *6th International Conference on Human System Interactions*, Sopot, Poland, pp. 210-213, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[61] Silvia Monica Feraru, Dagmar Schuller, and Bj÷rn Schuller, "Cross-Language Acoustic Emotion Recognition: An Overviewand Some Tendencies," *International Conference on Affective Computing and Intelligent Interaction*, Xi'an, China, pp. 125-131, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[62] Iker Luengo, Eva Navas, and Inmaculada Hernáez, "Feature Analysis and Evaluation for Automatic Emotion Identification in Speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490-501, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[63] André Stuhlsatz et al., "Deep Neural Networks for Acoustic Emotion Recognition: Raising the Benchmarks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, pp. 5688-5691, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[64] Wang Yutai et al., "Speaker Recognition Based on Dynamic MFCC Parameters," *International Conference on Image Analysis and Signal Processing*, Linhai, China, pp. 406-409, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[65] Nan Ding et al., "Speech Emotion Features Selection Based on BBO-SVM," *Tenth International Conference on Advanced Computational Intelligence*, Xiamen, China, pp. 210-216, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[66] Yaxin Sun, Guihua Wen, and Jiabing Wang, "Weighted Spectral Features Based on Local Hu Moments for Speech Emotion Recognition," *Biomedical Signal Processing and Control*, vol. 18, pp. 80-90, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[67] Maxim Sidorov et al., "Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm," *International Conference on Language Resources and Evaluation*, Reykjavik, Iceland, pp. 3481-3485, 2014. [Google Scholar] [Publisher Link]

[68] Anton Batliner et al., "Combining Efforts for Improving Automatic Classification of Emotional User States," *Procedding 5th Slovenian 1st International Language Technology Conference*, pp. 240-245, 2006. [Google Scholar] [Publisher Link]

[69] Xinzhou Xu et al., "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 795-808, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[70] B. Yang, and M. Lugger, "Emotion Recognition from Speech Signals using New Harmony Features," *Signal Processing*, vol. 90, no. 5, pp. 1415-1423, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[71] Anuja Bombatkar et al., "Emotion Recognition Using Speech Processing Using K-Nearest Neighbor Algorithm," *International Journal of Engineering Research and Applications*, pp. 68-71, 2014. [Google Scholar] [Publisher Link]

[72] Md. Kamruzzaman Sarker, Kazi Md. Rokibul Alam, and Md. Arifuzzaman, "Emotion Recognition from Speech based on Relevant Feature and Majority Voting," *International Conference on Informatics, Electronics & Vision*, Dhaka, Bangladesh, pp. 1-5, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[73] Shulan Xia et al., "An Improved Algorithm of Speech EmotionRecognition," *International Journal of u- and e- Service, Science and Technology*, vol. 8, no. 12, pp. 217-226, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[74] Andrzej Majkowski et al., "Classification of Emotions from Speech Signal," *Signal Processing: Algorithms, Architectures, Arrangements, and Applications*, Poznan, Poland, pp. 276-281, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[75] Yan Wang, and Weiping Hu, "Speech Emotion Recognition Based on Improved MFCC," *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, Hohhot China, pp. 1-7, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[76] Roy Patterson et al., "An Efficient Auditory Filterbank Based on the Gammatone Function," *Speech-Group Meeting of the Institute of Acoustics on Auditory Modelling*, pp. 1-33, 1987. [Google Scholar]

[77] X. Yang, K. Wang, and S.A. Shamma, "Auditory Representations of Acoustic Signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824 -839, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[78] Malcolm Slaney, "*An Efficient Implementation of the Patterson-Holdsworth Auditory Filterbank*," Apple Computer Technical Report, no. 35, 1993. [Google Scholar]

[79] Ludger Solbach, Rolf Wöhrmann, and Jörg Kliewer, *The Complex-Valued Continuous Wavelet Transforms as a Preprocessor for Auditory Scene Analysis*, 1st ed., Computational Auditory Scene Analysis, CRC Press, pp. 273-292, 1998. [Google Scholar] [Publisher Link]

[80] Stéphane G. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, pp. 1-577, 1998. [Google Scholar] [Publisher Link]

[81] Arun Venkitaraman, Aniruddha Adiga, and Chandra Sekhar Seelamantula, "Auditory Motivated Gammatone Wavelet Transform," *Signal Processing*, vol. 94, pp. 608-619, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[82] Gerald R. Patterson, John B. Reid, and Thomas J. Dishion, *Antisocial Boys*, Castalia Publishing Company, pp. 1-193, 1992. [Google Scholar] [Publisher Link]

[83] Erika Hoff-Ginsberg et al., " Maternal Speech and the Child's Development of Syntax: A Further Look," *Journal of Child Language*, vol. 17, no. 1, pp. 85-99, 1990. [CrossRef] [Google Scholar] [Publisher Link]

[84] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer Science & Business Media, pp. 1-700, 2012. [Google Scholar] [Publisher Link]

[85] L. Rabiner et al., "A Comparative Performance Study of Several Pitch Detection Algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399-418, 1976. [CrossRef] [Google Scholar] [Publisher Link]

[86] Sira Gonzalez, and Mike Brookes, "A Pitch Estimation Filter Robust to High Levels of Noise (PEFAC)," *19th European Signal Processing Conference*, Barcelona, Spain, pp. 451-455, 2011. [Google Scholar] [Publisher Link]

[87] Denis Byrne et al., "An International Comparison of Long-Term Average Speech Spectra," *The Journal of the Acoustical Society of America*, vol. 96, pp. 2108-2120, 1994. [CrossRef] [Google Scholar] [Publisher Link]

[88] Mike Brookes, VOICEBOX: A Speech Processing Toolbox for MATLAB, 1997. [Google Scholar] [Publisher Link]

[89] Ross Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580-587, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[90] Swati Pahune, and Nilu Mishra, "Emotion Recognition through Combination of Speech and Image Processing," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 2, pp. 134-137, 2015. [Google Scholar] [Publisher Link]

[91] Vishal B. Waghmare et al., "Development of Isolated Marathi Words Emotional Speech Database," *International Journal of Computer Applications*, vol. 94, no. 4, pp. 19-22, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[92] Vishal B. Waghmare et al., "Emotion Recognition System from Artificial Marathi Speech Using MFCC and LDA Techniques," *Fifth International Conference on Advances in Communication, Network, and Computing*, pp. 1-9, 2014. [Google Scholar]

[93] Wei Jiang et al., "Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network," *Sensors*, vol. 19, no. 12, pp. 1-15, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[94] Peng Song et al., "Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 265-275, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[95] YeSim Ülgen Sonmez, and Asaf Varol, "New Trends in Speech Emotion Recognition," *7th International Symposium on Digital Forensics and Security*, Barcelos, Portugal, pp. 1-7, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[96] Jianyou Wang et al., " Speech Emotion Recognition with Dual-Sequence LSTM Architecture," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, Barcelona, Spain, pp. 6474-6478, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[97] Noushin Hajarolasvadi, and Hasan Demirel, "3D CNN-Based Speech Emotion Recognition Using K- Means Clustering and Spectrograms," *Entropy*, vol. 21, no. 5, pp. 1-17, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[98] Abdul Malik Badshah et al., "Deep Features-Based Speech Emotion Recognition for Smart Affective Services," *Multimedia Tools and Applications*, vol. 78, pp. 5571-5589, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[99] Wootaek Lim, Daeyoung Jang, and Taejin Lee, "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Jeju, Korea (South), pp. 1-4, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[100] Zhengwei Huang et al., "Speech Emotion Recognition Using CNN," *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, Florida, USA, pp. 801-804, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[101] Ping Zhou et al., "Speech Emotion Recognition Based on MixedMFCC," *Applied Mechanics and Materials*, vol. 249-250, pp. 1252-1258, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[102] Dalibor Mitrović, Matthias Zeppelzauer, and Christian Breiteneder, " Features for Content-Based Audio Retrieval," *Advances in Computers*, vol. 78, pp. 71-150, 2010. [CrossRef] [Google Scholar] [Publisher Link]