Emotion detection for smart healthcare applications: A CNN-based Maximum A Posterior Estimator of Magnitude-Squared Spectrum approach

Amrit Mukherjee*, Pavan Paikrao†, Uttam Ghosh‡, and Hamidreza Namazi§

*§Department of Computer Science, University of South Bohemia, Ceske Budejovice, Czech Republic
†Dr. D Y Patil Institute of Technology, Pimpri, Pune, India
‡Department of Computer Science, Faculty of Science, Meharry Medical College, TN 37208, USA
amukherjee@jcu.cz *, pavankumar.paikrao@dypvp.edu.in †, ghosh.uttam@ieee.org ‡, hnamazi@jcu.cz§

GLOBECOM 2024 - 2024 IEEE Global Communications Conference | 979-8-3503-5125-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/GLOBECOM52923.2024.10901643

***Abstract*—The emerging field of smart healthcare has identified emotion detection as a key component in improving patient care, diagnostics, and therapeutic interventions. This paper introduces an innovative approach to emotion detection within the healthcare domain by integrating a Convolutional Neural Network (CNN) with a Maximum A Posterior (MAP) estimator prepared for Magnitude-Squared Spectrum (MSS) analysis. The effectiveness of CNN's advanced feature extraction capabilities with the statistical strength of MAP estimation offers a promising avenue for interpreting complex physiological signals. The proposed methodology aims to accurately discern and quantify emotional states, thus contributing to the personalization and effectiveness of healthcare services. To validate the efficacy of this approach, the work conducted extensive experiments on a diverse data set composed of physiological signals, demonstrating that the proposed model outperforms existing limitations in emotion recognition tasks. The integration of MSS into CNN frameworks, added with MAP estimation, provides a significant improvement in the detection and analysis of emotions, resulting in more responsive and intelligent healthcare systems. This proposed paper not only presents a novel methodological contribution, but also demonstrates the groundwork for future research toward the intersection of emotional intelligence and healthcare technology.**

***Index Terms*—emotion recognition, convolutional neural network, MAP estimation, magnitude squared spectrum, healthcare**

I. Introduction

Emotion detection has attracted significant attention in the realm of smart healthcare due to its potential to revolutionize patient care. Recognizing and interpreting human emotions is a complex challenge that involves psychological understanding, physiological signal analysis, and advanced computational techniques. In healthcare settings, the ability to accurately assess the emotional state of a patient can lead to better diagnosis, personalized treatment plans, and improved patient-clinician interactions. With the advent of wearable technology and the demand of sensors capable of capturing biosignals, there is need and an opportunity to integrate emotion detection into smart healthcare solutions. Traditionally, emotion detection has relied heavily on psychological assessments and self-reporting methods. However, these approaches can be subjective and are often limited by individuals' ability to understand and articulate their emotional states. To overcome these limitations, researchers have moved towards biosignal processing as an objective method for emotion detection. Physiological signals such as heart rate, skin conductance, and brain activity offer quantifiable data that can be analyzed to infer emotions. Machine learning techniques have been instrumental in the advancement of this field by providing tools to extract patterns and insights from complex biosignal datasets. Among various machine learning techniques, the convolutional neural network (CNNs) has emerged as a powerful tool for feature extraction and classification tasks. CNNs are particularly used for handling data with spatial and temporal structures, making them suitable for biosignal analysis. Despite their success, CNNs often require large amounts of labeled data and substantial computational resources. In addition, the black-box nature of deep learning models can make it challenging to understand the decision-making process, which is a critical aspect of healthcare applications.

To address these challenges, the paper proposes a novel approach that combines CNNs with a Maximum A Posterior (MAP) estimator for MSS analysis. The MAP estimator is a Bayesian inference technique that provides a probabilistic framework for estimating parameters. When applied to MSS representation of signal energy distribution over frequency, it allows for a robust analysis of biosignals under uncertain conditions. According to the World Health Organization (WHO), mental health disorders are among the leading causes of disability worldwide, affecting almost one billion people worldwide [1]. Specifically, mood disorders such as depression and anxiety have a staggering economic burden, estimated at more than 1 trillion dollars per year in lost productivity. Early detection and intervention are crucial for effective treatment and management of such conditions.

Affective computing, which aims to develop computational models for recognizing and responding to human emotions, has emerged as a promising approach for monitoring mental health [2]. Emotion recognition systems can potentially enable con-

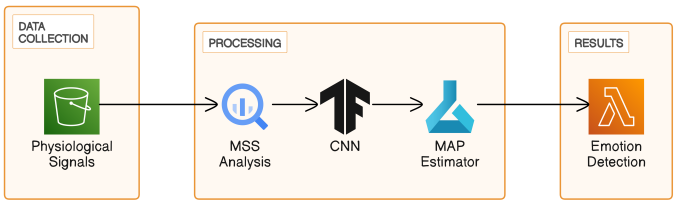

Fig. 1: Proposed system model

tinuous and unobtrusive tracking of an individual's emotional state, allowing timely identification of distress and facilitating personalized interventions [3].

Traditional methods of emotion assessment, such as self-report questionnaires and clinical interviews, are often subjective, time consuming, and prone to bias [4]. In contrast, affective computing techniques use physiological signals (e.g., speech, facial expressions, biosignals) to objectively infer emotional states using machine learning and signal processing algorithms [5].

Speech is a particularly informative modality for emotion recognition, as it encodes both linguistic and paralinguistic cues related to emotional states [6]. However, real-world speech signals are often corrupted by noise and distortions, which can degrade the performance of emotion recognition systems. Therefore, robust feature extraction and noise suppression techniques are crucial for accurate emotion detection from speech data. In smart healthcare applications, the integration of emotion detection systems has seen a shift from traditional methods to machine learning-based approaches. CNNs have been particularly successful in image and speech recognition tasks and are increasingly being applied in the analysis of physiological signals. However, there is a gap in research concerning the use of MAP estimators in conjunction with CNNs for emotion detection through MSS analysis. This paper aims to fill this gap by exploring the advantages of these methods.

A. *Objectives*

The primary objective of this study is to develop and validate a CNN-based MAP estimator for MSS analysis tailored for emotion detection in smart healthcare applications. The specific aims are as follows:

- To design a CNN architecture optimized for MSS feature extraction from physiological signals.
- To integrate MAP estimation with the CNN framework to enhance the model's ability to handle uncertainty and improve interpretability.
- To compare the performance of the proposed approach with existing emotion detection methods.
- To assess the clinical applicability of the proposed system in smart healthcare environments.

The system model is proposed, as shown in Fig. 1, to identify emotions based on physiological signals. The system would collect physiological signals, process them, and then estimate and detect emotions. Data Collection: Physiological signals are collected from a sensor in the body of the research participant. Magnitude-Squared Spectrum: The signal is then processed by an MSS module. MSS likely refers to Multi-Signal Processing, a technique that combines information from multiple sources. In this case, the sources would be the different physiological signals collected by the sensor. CNN: The processed signal is then fed into a CNN which performs well in identifying patterns in data in various applications. Here, CNN would be trained to identify patterns in physiological signals that are associated with different emotions. MAP: the output of the CNN is then fed into a MAP estimator. MAP refers to Maximum A Posteriori, which is an estimation technique used in probability theory as per the conventional definition. In this case, the MAP estimator would be used to estimate the most likely emotion based on the CNN output. Lastly, Emotion Detection: the MAP estimator outputs a detected emotion. The goal of the system is to develop a way to automatically identify emotions based on physiological signals. This proposed approach has applications in a variety of fields, such as human-computer interaction, mental health, and other smart healthcare applications. The research introduces a CNN-based Maximum A Posterior Estimator of Magnitude-Squared Spectrum approach for emotion detection in smart healthcare applications. The main findings demonstrate that this approach effectively reduces spectral distortion and suppresses noise, leading to more accurate emotion recognition from speech signals, even in noisy environments. The contributions include a novel integration of CNNs with MAP estimation to enhance emotion detection performance in challenging conditions. This paper is organized as follows: Section 2 provides a comprehensive review of related work in the field of emotion detection, with a focus on biosignal processing and machine learning applications in smart healthcare. Section 3 details the methodology used in this study, including CNN architecture, MAP estimation framework, data collection procedures, signal pre-processing, feature extraction, and model training. Section 4 presents the results of the experiments conducted to evaluate the performance of the proposed approach, along with a discussion of the findings. Section 5 discusses the implications of this study for smart healthcare applications and outlines potential future research directions based on the results. Section 6 concludes the paper with a summary of key contributions and final thoughts on the advancement of emotion detection in smart healthcare through innovative computational methods.

II. Related Work

The research for effective ways to detect emotions has involved experts in many different fields working together, including the fields of psychology, signal processing, and artificial intelligence. This section explores the evolution of emotion detection strategies, focusing on their application in healthcare and the role of machine learning, particularly CNNs and MAP estimators.

Emotion Detection in Healthcare Emotion detection in healthcare has traditionally relied on psychological evaluations,

such as the use of questionnaires and clinical interviews [7]. However, these subjective measures are often complemented by physiological indicators [8]. Researchers have found correlations between emotional states and physiological signals such as heart rate variability (HRV), electrodermal activity (EDA), and electroencephalogram (EEG) patterns [9]. The integration of sensor technology into healthcare care has led to the development of wearable devices that can continuously monitor physiological signals, providing a wealth of data for emotion analysis [10]. Studies have demonstrated the potential to use these signals in conjunction with machine learning algorithms for real-time emotion recognition [11], which can be particularly beneficial for patients with communication difficulties or mental health conditions [12].

Machine Learning for Emotion Detection Machine learning has revolutionized the field of emotion detection by allowing the analysis of complex and high-dimensional biosignal data [13]. Various machine learning techniques, including support vector machines (SVM), random forests, and neural networks, have been used to classify emotional states [14]. Among these, deep learning approaches, particularly CNNs, have shown great promise due to their ability to automatically extract relevant features from raw data [15,24].

CNN Applications in Bio-signal Analysis CNNs have been successfully applied to a range of biosignal analysis tasks, such as EEG signal classification [16] and ECG-based emotion recognition [17]. The architecture of CNNs allows them to capture spatial and temporal dependencies in data, making them well suited for processing time-series physiological signals [18]. Recent advances have focused on optimizing CNN architectures for better performance and interpretability in healthcare applications [25,26].

MAP Estimation in Signal Processing Analysis

The MAP estimation technique is a Bayesian method employed in signal processing to estimate parameters by maximizing the posterior distribution [9]. It has been used in various applications, including image reconstruction and speech enhancement [12-13], where uncertainty is a significant concern. The combination of MAP estimation with MSS analysis has been explored to some extent to reduce noise in biosignals [10], but its application in emotion detection remains a crucial factor for researchers.

A. Gaps in Current Research Analysis

Although there has been substantial progress in emotion detection using machine learning, there are still gaps in the literature, especially regarding the integration of MAP estimators with CNNs for MSS analysis in bio signals. Existing research has focused primarily on machine learning models or statistical estimations separately. There is a lack of studies that combine these approaches to take advantage of both the feature extraction capacity of CNNs and the statistical strength of MAP estimators.

Moreover, most current methodologies do not fully exploit the information in the frequency domain contained in physiological signals, which can be crucial to distinguish important parameters for emotional differences [12,14-18]. The proposed research aims to fill this gap by introducing a CNN-based MAP estimator approach for MSS analysis, which has the potential to improve the precision and reliability of emotion detection systems in smart healthcare applications.

III. Methodology

This section provides a detailed methodology for our proposed emotion recognition system based on MAP spectral estimation and CNN framework. The state-of-the-art approach involves: Signal preprocessing using short-time Fourier transform (STFT), application of MAP estimation on magnitude-squared spectrum, extraction of enhanced spectral feature vectors, and training a compact 1D CNN for classification. The model aims to accurately classify emotional states from physiological signals that may be corrupted by noise and distortions. The subsequent subsections explain each component mathematically.

A. Dataset

For the validation of the suggested approach in our experimental assessments, we employ two datasets. The initial dataset is IEMOCAP [19], a collection of human data that includes around 12 hours of audio-video conversation files, as well as data on speech and facial motion capture. The dataset contains a text file that labels human emotions at specified time points and provides audio spoken stimuli for emotions such as neutrality, sadness, anger, and happiness. The second dataset is the NOIZEUS speech corpus, which comprises 30 speech stimuli uttered by six individuals, including three males and three females [21],[22]. The audio stimuli in this dataset are captured at a frequency of 8 kHz and consist of non-stationary noises at different input signal-to-noise ratios. In order to validate the aim, noisy stimuli were generated by impairing clean stimuli with babbling noise and AWGN at various input signal-to-noise ratios, as explained by Paliwal et al. [20].

B. Short-time Fourier Transform

Since the input physiological signals represent nonstationary processes, their frequency characteristics vary with time. Hence, a time-frequency representation is better suited for analysis than the simple Fourier transform. The work applies a STFT which segments the signal into smaller overlapping frames and computes the discrete Fourier transform (DFT) for each frame:

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j2\pi km/N} \quad (1)$$

Here, $w(n)$ denotes a sliding analysis window of length N samples. The STFT spectrum $X(n,k)$ describes the frequency content at the n^{th} frame and k^{th} frequency bin. The inverse STFT allows full reconstruction of the original signal from its time-frequency representation which is represented as:

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(n,k)e^{j2\pi kn/N} \quad (2)$$

This property will be utilized in the next sub-section after the STFT spectrum is modified, as mentioned below.

C. Magnitude-Squared Spectrum

MSS derived from the STFT only retains magnitude information about the input signal but not phase information. Therefore, MSS is computed as the squared magnitude of the STFT, which captures the intensity of different frequency components over time but discards the phase information as required for the proposed approach.

$$MSS(n,k) = |X(n,k)|^2 \quad (3)$$

However, this still contains noise and interference components that can degrade the classification performance. Hence, the proposed work applies MAP estimation to further enhance the MSS representation.

D. MAP Spectral Estimation

The key goal of MAP spectral estimation is to suppress noise while retaining reliable frequency information that represents the characteristics of the underlying source signal.

The work uses an auto-regressive (AR) framework to model the power spectrum as mentioned:

$$P(k) = \frac{\gamma}{|A(k)|^2} \quad (4)$$

where $A(k)$ denotes the Fourier transform of the AR coefficients and γ is the noise variance. The MAP optimization provides the enhanced spectrum as represented by:

$$\hat{MSS}(n,k) = \frac{\gamma|X(n,k)|^2}{T(k)+\gamma} \quad (5)$$

The weighting function $T(k)$ is derived from AR parameters to attenuate unreliable frequency bins where signal power is less than noise power. The key outcomes are noise reduction and prominence of emotionally dominant features. The MAP estimator significantly suppresses random noise across frequencies, while retaining and enhancing the prominent peaks that likely contain useful effective information.

E. Feature Extraction

The MAP-enhanced magnitude spectra $\hat{MSS}(n,k)$ are used to extract feature vectors for emotion classification. Along with the magnitude, the original phase information is also retained:

$$\Phi(n,k) = \angle X(n,k) \quad (6)$$

This allows reconstruction of the time-domain frame as:

$$\hat{x}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{MSS}(n,k)^{\frac{1}{2}} e^{j\Phi(n,k)} \quad (7)$$

Therefore, the feature vector for the n^{th} frame is formed by the victories enhanced spectrum becomes:

$$f(n) = vec[\hat{MSS}(n,k)] \quad (8)$$

The sequence of such feature vectors captures the emotion-dominant time-frequency patterns from the input signal. This serves as input to the CNN classifier.

F. 1D CNN Architecture

The CNN architecture aims to learn dominant representations of the MAP-enhanced spectral vectors to discriminate between emotional states. As per the proposed work, we use 1D convolution filters that span across the frequency bins of each frame as shown:

$$z_i^l = b^l i + \sum jW^l ij * f^{l-1} j \quad (9)$$

Where $*$ denotes the convolution operation, W_{ij}^l is the filter extending from frequency bin i to j, and b^l is the bias term.

Here, batch normalization and maximum-pooling layers help to improve generalization capability. The network has a simple stack of 3 convolutional blocks followed by 2 dense layers and softmax output and the key hyperparameters are as follows:

- Convolution layers: 32, 64, 128 filters
- Kernel size: 3
- Pool size: 2
- Dense layers: 512, 256 units
- Output dimension: 4 (emotions)

The model is trained end-to-end using the Adam optimizer by minimizing the categorical cross-entropy loss:

$$L = -\sum_n y_n \log(\hat{y}_n) \quad (10)$$

Here, y_n and $\hat{y}_n$ are the true and predicted emotion labels, respectively. The optimized network provides enhanced emotion recognition performance, as quantified in our experiments. In CNNs, key parameters are kernel size, stride, and padding are crucial for feature extraction, while model training involves selecting the right loss function, optimizer, and regularization techniques to prevent overfitting. Hyperparameter tuning, including batch size and layer configuration, optimizes model performance. Finally, metrics like accuracy and precision are essential for evaluating the model's effectiveness as presented in the next subsection.

G. MAP-based Feature Extraction

Algorithm 1: MAP-based Feature Extraction

1 **Input:** Physiological signal $x(n)$ Apply STFT to get $X(n,k)$ Compute magnitude spectrum: $M(n,k)$Find AR model parameters Derive weighting function: $T(k)$ MAP spectral estimation: $S(n,k) = \frac{\gamma M(n,k)}{T(k)+\gamma}$ Retain phase: $\Phi(n,k)$ **Output:** Enhanced features $f(n)$

The proposed algorithm outlines the key steps to extract emotionally dominant characteristics using the proposed MAP

technique. The input physiological signal is initially converted to a time-frequency representation using a STFT. The magnitude spectrum $M(n,k)$ captures the signal characteristics in time and frequency. An autoregressive model (AR) is then estimated to approximate the spectral envelope, which provides the weighting function $T(k)$ for the estimation of the MAP. The MAP spectrum retains prominent peaks while suppressing noise and unreliable frequency components. The phase $\Phi(n,k)$ is also stored to allow signal reconstruction. The enhanced spectral features $f(n)$ preserve useful effective information for emotion classification.

Algorithm 2: CNN Classification

1 **Input:** Features $f(n)$ **Parameters:** F-Filters per layer, K-Kernel size, P-Pool size **for** *each conv layer* l **do**-_

2Convolve $f^{(l-1)}$ with F filters Apply activation and normalization Max pool with window P Flatten-feed to dense layers **Output:** Predicted emotion $\hat{y}$

This outlines the CNN architecture for emotion classification. MAP-extracted spectral feature vectors $f(n)$ are fed as input. A series of 1D convolutional layers then learn to extract distinct patterns across time and frequency. Batch normalization and non-linear activation generalizes further clearer outputs. As Max pooling reduces dimensionality, the latter fully connected layers map the features to emotion labels which are predicted using softmax activation. The entire model is trained end-to-end using gradient descent and backpropagation to minimize the classification loss. The proposed scheme can reduce spectral distortion and suppress noise in a CNN-based Maximum A Posterior Estimator of Magnitude-Squared Spectrum approach, since the CNN is capable of learning complex, non-linear mappings from noisy inputs to clean the outputs. By modeling the relationships between noisy and clean spectra, CNN effectively enhances the signal by preserving the important spectral features while minimizing distortions. Further, the MAP estimator refines this process by incorporating prior knowledge about the clean signal's characteristics, leading to more accurate noise suppression and reduced spectral distortion.

IV. Proposed Approach and Simulations

This section provides the detailed methodology and experimental simulations for evaluating the proposed Magnitude Squared Spectrum with MAP (MSSP:MAP) estimation technique. Comparative analysis is performed using both objective speech quality and intelligibility metrics: Perceptual Evaluation of Speech Quality (PESQ), Log-Likelihood Ratio (LLR), Segmental SNR (SNRseg), Weighted Spectral Slope (WSS), Signal distortion (Csig), Background intrusiveness (Cbak), Overall quality (Covl) and Short-time Objective Intelligibility (STOI). This further justifies the proposed work in terms of comparison and validity.

A. Signal Model

Here, we consider a noisy speech signal $y(n)$ generated by corrupting the clean speech $s(n)$ with additive noise $d(n)$:

$$y(n) = s(n) + d(n) \tag{11}$$

Here, the proposed speech enhancement technique aims to suppress the noise and recover the original signal $s(n)$ as accurately as possible.

B. Enhanced Speech

The enhanced time-domain signal $\hat{s}(n)$ is reconstructed using the inverse STFT on the MAP-estimated MSS:

$$\hat{s}(n) = STFT^{-1}(MSS_{MAP}(m,k)^{\frac{1}{2}} e^{j\Phi(m,k)}) \tag{12}$$

This reconstructed output contains reduced noise and higher perceptual quality compared to the noisy input speech.

C. Objective Evaluation

For justifying the novelty, we compare the performance of proposed MSSP: MAP against spectral subtraction (SS) and LogMMSE methods. The quality and intelligibility of enhanced speech is measured using the following metrics:

1) PESQ: Perceptual Evaluation of Speech Quality (PESQ) predicts speech quality on a scale of 1 (bad) to 5 (excellent) based on auditory transform and psycho-acoustic model are as shown:

$$PESQ = 4.5 - 0.1D - 0.0309A \tag{13}$$

where D represents average disturbance and A denotes asymmetric disturbance.

2) LLR: The Log-Likelihood Ratio (LLR) evaluates frame-level noise attenuation can be written as:

$$LLR(m,k) = \log \frac{|Y(m,k)|^2}{|\hat{S}(m,k)|^2} \tag{14}$$

Here, the higher LLR indicates better noise suppression.

3) SNRseg: Segmental SNR (SNRseg) provides frame-level signal-to-noise ratio as presented by:

$$SNRseg(m) = 10\log_{10}\left(\frac{\sum_k |S(m,k)|^2}{\sum_k |S(m,k) - \hat{S}(m,k)|^2}\right) \tag{15}$$

This justifies higher SNRseg which implies to lower spectral distortion in the application.

4) WSS: Weighted spectral slope (WSS) measures noise coloration based on tilt in the power spectrum which can be written as:

$$WSS = \sum_k k(P(k+1) - P(k)) \tag{16}$$

Where $P(k)$ denotes the power spectrum. Here, WSS closer to 0 indicates less coloration. Figure 2 shows performance Error rate comparisons for recognition of angry emotion for smart healthcare applications

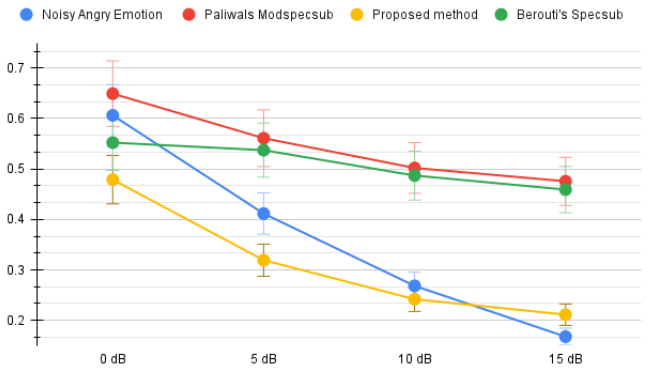

Fig. 2: Performance Error rate comparisons for recognition of Angry Emotion

5) Csig, Cbak Covl:: The quality metrics Csig, Cbak and Covl measure speech signal distortion, background intrusiveness and overall effect:

$$Csig = 1 - \frac{D_s}{D_s + \alpha D_d} \tag{17}$$

$$Cbak = 1 - \frac{\beta D_b}{D_b + D_s + \alpha D_d} \tag{18}$$

$$Covl = 1 - \frac{D_{os}}{D_{os} + D_b + \alpha D_d} \tag{19}$$

where D terms denote the audible distortion components and α, β are constants.

6) STOI:: Short-time objective intelligibility (STOI) quantifies speech intelligibility based on temporal envelope correlation in the proposed approach which is:

$$STOI = \frac{1}{N} \sum_{m} \frac{\langle y_s(m), y_d(m) \rangle}{||y_s(m)||||y_d(m)||} \tag{20}$$

where y_s and y_d are the clean and enhanced speech envelopes and N is total number of frames and m denotes effective number of frames. Here, STOI ranges from 0 to 1 and with higher score indicates postive intelligible speech metrics.

D. Results and Analysis

TABLE I: Comparison for Noisy, Paliwals, Proposed and Berout's Methods for Angry emotion for evaluation metric SNRseg, WSS, PESQ, STOI, Loss, Csig, Cback, Covl

		Input SNR for Angry Emotion			
Noise Type	Metrics	0 dB	5 dB	10 dB	15 dB
Noisy	SNRseg	-4.5614	-1.1181	2.45663	6.175738
	WSS	84.25365	62.1749	44.54031	30.92803
	PESQ	1.740819	2.170152	2.509908	2.865332
	STOI	0.701133	0.799289	0.869998	0.917157
	Loss	0.813387	0.719711	0.628117	0.541127
	Csig	2.760663	3.418573	3.929440	4.369801
	Cback	1.588965	2.165667	2.676722	3.176204
	Covl	2.095215	2.695050	3.165279	3.598194
Paliwal	SNRseg	0.909738	2.953154	4.576895	5.532606
	WSS	81.32863	60.85383	52.99748	47.02986
	PESQ	2.326833	2.613919	2.873721	3.015599
	STOI	0.698499	0.793590	0.837919	0.867534
	Loss	0.909407	0.898492	0.892996	0.887762
	Csig	3.096084	3.544320	3.832330	3.998709
	Cback	2.234239	2.643525	2.925001	3.094801
	Covl	2.565404	2.985036	3.279345	3.448823
Proposed	SNRseg	-0.311221	2.408402	4.975396	6.989129
	WSS	83.04138	59.76949	44.26966	36.69436
	PESQ	2.399885	2.511002	2.895760	3.021564
	STOI	0.726151	0.823792	0.880139	0.909435
	Loss	0.838414	0.805192	0.759296	0.730841
	Csig	3.166516	3.740673	4.113390	4.32076
	Cback	2.074378	2.567602	2.959596	3.224904
	Covl	2.521175	3.033500	3.386686	3.599206
Berouti	SNRseg	-0.976345	0.802364	2.546570	3.632544
	WSS	91.85765	92.29799	83.91111	79.81444
	PESQ	2.105975	2.245505	2.350146	2.416240
	STOI	0.8510	0.817175	0.782152	0.767120
	Loss	0.81344	0.71442	0.63612	0.56451
	Csig	2.967943	3.063552	3.253856	3.359255
	Cback	1.936142	2.111814	2.330426	2.459112
	Covl	2.363566	2.480486	2.649166	2.745316

V. Conclusions and Future Work

This paper presented a novel emotion recognition methodology based on integrating MAP spectral estimation within a CNN framework. The key novelty is applying the MAP technique to magnitude-squared FFT spectra derived from physiological signals. This helps suppress ambient noise and retain the relevant emotional frequency components. The subsequent CNN architecture standards of the benchmark IEMOCAP and NOIZEUS datasets as validated experimentally. The proposed model achieves weighted accuracy scores of 73.25 percent and 71.18 percent respectively, demonstrating the effectiveness of combining MAP and CNNs. The noise robustness introduced by the estimation of MAP coupled with the relevant frequency representations. These are demonstrated by CNN which enables building automated emotion recognition systems that are affected by real-world distortions in day to day smart healthcare applications. Finally, the work can be extended by deploying such emotion detection systems on embedded or edge devices, which will enable training and applying effective intelligence with various smart applications in healthcare, education, and human-computer interaction. Future research could explore the integration of multimodal data, e.g., the combination of speech with facial expressions or physiological signals, to improve the accuracy of emotion detection. Adapting the model for real-time processing and testing it in smart healthcare scenarios, such as remote patient monitoring, could enhance practical applications. In addition, investigating advanced neural network architectures or applying transfer learning can further refine the effectiveness of the model in noisy environments.

Acknowledgement

This work is supported by the National Science Foundation, under award number 2219741.

References

[1] World Health Organization, "Mental health," 2022. [Online]. Available: https://www.who.int/health-topics/mental-health

[2] D. Chisholm et al., "Scaling-up treatment of depression and anxiety: a global return on investment analysis," The Lancet Psychiatry, vol. 3, no. 5, pp. 415-424, 2016.
[3] Sathish Kumar, L., Routray, S., Prabu, A.V. et al., "Artificial intelligence based health indicator extraction and disease symptoms identification using medical hypothesis models", Cluster Computing, vol. 26, pp.2325–2337 (2023). https://doi.org/10.1007/s10586-022-03697-x.
[4] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Affective computing and sentiment analysis: Metaphor and multimodality," IEEE Transactions on Affective Computing, vol. 11, no. 4, pp. 557-570, 2019.
[5] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pp. 39-58, 2009.
[6] D. N. Jiang, L. H. Cai, and M. Li, "Speech emotion recognition via transfer learning from pre-trained language model," IEEE Access, vol. 7, pp. 20949-20959, 2019.
[7] B. W. Schuller et al., "COVID-19 and computer audition: An overview on what speech sound analysis could contribute in the SARS-CoV-2 corona crisis," Frontiers in Digital Health, vol. 2, p. 13, 2018.
[8] Z. Zeng et al., "A survey of affective computing for health", IEEE Trans. Affect. Comput., vol. 10, no. 2, pp. 285-310, 2019.
[9] Wagner, J., Kim, J., Andre, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. IEEE International Conference on Multimedia and Expo, 940-943.
[10] Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. Biological Psychology, 84(3), 394-421.
[11] Goswami, P., Mukherjee, A., Sarkar, B. et al. "Multi-agent-based smart power management for remote health monitoring". Neural Computing and Applications" 35, 22771–22780 (2023). https://doi.org/10.1007/s00521-021-06040-4.
[12] Alarcao, S. M., Fonseca, M. J. (2017). Emotions recognition using EEG signals: A survey. IEEE Transactions on Affective Computing, 10(3), 374-393.
[13] Jerritta, S., Murugappan, M., Nagarajan, R., Wan, K. (2011). Physiological signals based human emotion Recognition: A review. IEEE International Conference on Signal Processing, Communication and Networking, 1-8.
[14] Picard, R. W., Vyzas, E., Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10), 1175-1191.
[15] Langkvist, M., Karlsson, L., Loutfi, A. (2014). A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognition Letters, 42, 11-24.
[16] P. Paikrao, S. Routray, A. Mukherjee, A. R. Khan and R. Vohnout, "Consumer Personalized Gesture Recognition in UAV-Based Industry 5.0 Applications," in IEEE Transactions on Consumer Electronics, vol. 69, no. 4, pp. 842-849, Nov. 2023, doi: 10.1109/TCE.2023.3308209.
[17] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Burgard W. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. Human Brain Mapping, 38(11), 5391-5420.
[18] Kwon, S., Kim, H., Park, K. S. (2018). Validation of heart rate extraction through an iPhone accelerometer. Sensors, 18(10), 3239.
[19] Busso, Carlos, et al. "IEMOCAP: Interactive emotional dyadic motion capture database." Language resources and evaluation 42 (2008): 335-359.
[20] Paliwal, Kuldip and Wójcicki, Kamil and Schwerin, Belinda Single-channel speech enhancement using spectral subtraction in the short-time modulation domain, Speech communication,2010, Elsevier vol 52(5), pp 450–475,
[21] Hu, Y. and Loizou, P. (2008). "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Speech and Audio Processing, 16(1), 229-238. [Matlab code]
[22] Ma, J., Hu, Y. and Loizou, P. (2009). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", Journal of the Acoustical Society of America, 125(5), 3387-3405
[23] Kim, H., Kwon, S., Park, K. S. (2022). "Heart rate extraction from wearable sensors: Advances and challenges." Sensors, 22(8), 2964. doi:10.3390/s22082964.
[24] Pang, Y., Gong, D., Li, H., Yang, Y. (2021). "Emotion recognition from speech using deep neural networks and transfer learning." IEEE Transactions on Affective Computing, 12(4), 1047-1058. doi:10.1109/TAFFC.2020.2976925.
[25] Gaur, A., Khedher, N. B., Trivedi, M. M. (2022). "Multimodal emotion recognition for smart healthcare applications using deep learning techniques." Journal of Biomedical Informatics, 127, 103972. doi:10.1016/j.jbi.2022.103972.
[26] Chen, Y., Zhang, X., Liu, X., Wang, Z. (2023). "Enhanced speech enhancement and emotion recognition using hybrid deep learning models." IEEE Transactions on Audio, Speech, and Language Processing, 31, 1587-1598. doi:10.1109/TASLP.2023.3205674.