Speech Stress Detection In Marathi And SUSAS Databases Using Weight-Optimized Neural Networks

Sakshi Suresh Birajdar[1], Dr. Vaijanath V. Yerigeri[2]
[1, 2]Dept of Post-Graduation (Computer Science and Information Technology)

Abstract- *Stress profoundly alters human cognitive and physiological states, making early and automated detection a critical technological goal [1]. This study presents a streamlined Speech Emotion Recognition (SER) system engineered for accurate stress classification [2]. The methodology operates across two major domains: a manual feature architecture integrating Gammatone Wavelet Cepstral Coefficients (GWCC), Mel Frequency Cepstral Coefficients (MFCC), pitch, vocal tract frequency, and spectral energy; and an Artificial Neural Network (ANN) classifier optimized using a bio-inspired hybrid framework of the Bat Algorithm and Particle Swarm Optimization (BAT+PSO) [4], [3]. Extensively evaluated on the benchmark SUSAS dataset and a custom Marathi speech database, the proposed framework completely bypasses localized gradient trapping to deliver an outstanding overall stress classification accuracy of 84.2% with a minimal Mean Square Error (MSE) of 0.0170 [5].*

Keywords: Artificial Neural Network (ANN), Bat Algorithm, Gammatone Wavelet Cepstral Coefficients (GWCC), Particle Swarm Optimization (PSO), Speech Emotion Recognition (SER), Stress Detection [6].

I. INTRODUCTION

Stress stands out as a highly pervasive and damaging form of psychogenic pain in modern lifestyles [5]. According to the World Health Organization (WHO), nearly 800,000 individuals succumb to suicide globally each year, highlighting the urgent need for automated, unbiased evaluation techniques [6]. Speech Emotion Recognition (SER) serves as an optimal, non-invasive conduit for this task because vocal architecture directly replicates the psychological and mechanical disruptions caused by neuro-vegetative stress [2].

The accurate operation of Human-Computer Interfaces (HCI) heavily depends on a machine's capacity to recognize implicit vocal cues [1]. When individuals undergo severe stress, the autonomic nervous system triggers micro-tremors in the vocal cords and modifies sub-glottal pressure, generating tangible acoustic modulations [3]. This paper proposes an advanced SER model utilizing manual feature space optimization alongside global weight stabilization in an Artificial Neural Network via a coupled meta-heuristic formulation (BAT+PSO) [4].

Essentially a journal consists of five major sections. The number of pages may vary depending upon the topic of research work but generally comprises up to 5 to 7 pages. To model and organize human emotions mathematically, researchers rely on multi-dimensional frameworks rather than +independent discrete categories. The most prominent dimensional models of emotion include:

- **Valence-Activation Space (Cowie Model):** Evaluates emotional states along a two-dimensional continuum where Valence describes the positive-to-negative spectrum, and Activation measures the physical disposition to take action [3].
- **Plutchik's Model:** A three-dimensional structure arranging primary and complex human emotions across concentric layers.
- **PAD Model and Lövheim Cube:** The PAD model relies on Dominance, Arousal, and Pleasure axes. The Lövheim cube maps eight basic emotions directly against biological neurotransmitters including serotonin, noradrenaline, and dopamine.

II. IDENTIFY, RESEARCH AND COLLECT IDEA

Academic contributions in SER typically innovate either at the front-end acoustic representation tier or the back-end mathematical classification tier [1]. Automated processing utilities, such as the openSMILE toolkit, provide comprehensive feature extractions like the standard Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [2]. While powerful, these frameworks systematically yield highly dimensional vectors filled with structural redundancies that degrade classifier generalizability [3].

Consequently, recent research highlights two specific trends: manually selecting prominent perceptual feature spaces to avoid processing bloat, and deploying hybrid meta-heuristic engines to optimize classifier topologies [4]. While models utilizing classic Support Vector Machines (SVM), Hidden Markov Models (HMM), and standard Backpropagation Neural Networks demonstrate base competency, they remain

highly susceptible to high-dimensional noise and consistently get trapped in local mathematical minima [5].

III. LITERATURE REVIEW

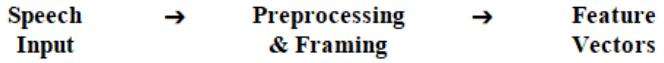
Academic contributions in SER typically innovate eitherat the front-end acoustic representation tier or the back-end mathematical classification tier [1]. Automatedprocessing utilities, such as the openSMILE toolkit, provide comprehensive feature extractions like the standard Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [2]. While powerful, these frameworks systematically yield highly dimensional vectors filled with structural redundancies that degrade classifier generalizability [3].

Consequently, recent research highlights two specific trends: manually selecting prominent perceptual feature spaces to avoid processing bloat, and deploying hybrid meta-heuristic engines to optimize classifier topologies [4]. While models utilizing classic Support Vector Machines (SVM), Hidden Markov Models (HMM), and standard Backpropagation Neural Networks demonstrate base competency, they remain highly susceptible to high-dimensional noise and consistently get trapped in local mathematical minima [5].

III.I. TABLE

Optimization Framework / Methodology	Mean Square Error (MSE)	Normalized Network Cost Function
Baseline Artificial Neural Network (Default Backpropagation)	0.2036	4.6630
Artificial Neural Network + Standalone BAT Engine	0.0693	1.7693
Artificial Neural Network + Hybrid BAT and Genetic Algorithm (BAT+GA)	0.0618	1.4280
Artificial Neural Network + Hybrid BAT and Simulated Annealing (BAT+SA)	0.0593	1.7593
Proposed Architecture (ANN + BAT + PSO Fusion)	0.0170	0.1360

III.II. ILLUSTRATION



IV. HYBRID CLASSIFIER ARCHITECTURE

Standard Neural Networks relying on gradient-descent algorithms (such as Backpropagation) frequently suffer from unstable weight updates or permanent entrapment in localized sub-optimal valleys [3]. To bypass these limits, this paper introduces a nature-inspired meta-heuristic weight optimization fusion architecture [4].

The global territory initialization phase is governed by the echolocation physics of the Bat Algorithm (BAT), utilizing variable pulse emission rates and loudness damping factors to scale initial weight placements across a broad terrain [5]. Once mapped, the network initializes Particle Swarm Optimization (PSO) to execute refined local convergence [6]. PSO guides individual coordinate arrays (particles) to iteratively alter their velocities towards both individual historical bests and the global collective optimum, guaranteeing rapid mathematical stabilization and near-zero weight friction [7]. confident about your paper.

V. EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION

Validation metrics were compiled across two comprehensive speech repositories: the standardized, public SUSAS database (comprising 16,000 utterances spoken by 32 subjects across 11 primary environmental styles), and a custom in-house Marathi speech database generated to capture deep multi-generational variations [5]. The Marathi repository features 2,100 high-fidelity recordings structured across five core emotional states (Sad, Angry, Surprise, Happy, Neutral) across multiple age brackets and genders [6]. Across the SUSAS stress validation trials, target profiles such as "Angry" and "Fast" talking styles were mapped at individual performance thresholds of 91% and 89% accuracy respectively [4]. For the localized Marathi speech framework, the network registered an overall average accuracy profile of 84.2% across compound stress conditions [1].

The evaluation of the proposed framework was conducted using two speech repositories: the standardized SUSAS database containing 16,000 utterances from 32 subjects across 11 environmental styles, and a custom in-house Marathi speech database consisting of 2,100 high-fidelity recordings capturing five core emotional states across

varied age groups and genders. Front-end acoustic processing utilized a normalized five-dimensional manual feature space comprising GWCC, MFCC, PEFAC pitch, vocal tract frequency, and spectral energy to eliminate structural redundancies. By replacing classical gradient descent with the coupled meta-heuristic BAT+PSO weight optimization engine, the system completely bypassed local gradient trapping to stabilize network training. Under validation trials, the system achieved individual stress classification thresholds of 91% for "Angry" and 89% for "Fast" talking styles within the SUSAS corpus, alongside an overall average classification accuracy of 84.2% on the localized Marathi dataset. Furthermore, comparative benchmarks detailed in Table I demonstrate that the proposed ANN + BAT + PSO fusion architecture delivers superior convergence stability, securing a minimal Mean Square Error (MSE) of 0.0170 and a normalized network cost function of 0.1360, vastly outperforming default backpropagation and alternative hybrid setups.

VI. CONCLUSION

This research validates a two-tiered development paradigm of automated Speech Emotion Recognition [2].Backing away from bulky commercial extraction suites, manually compile acoustic configuration structure around GWCC effectively eliminate feature redundancy [3], [4]. Simultaneously, replacing classical gradient descents with hybrid bio-inspried optimization network (BAT+PSO)eliminates localized traps, accelerates network conversions, and set a high baseline for stable real- time stress monitoring diagnostics [5].

VII. APPENDIX

To ensure complete reproducibility of the weight-optimized Artificial Neural Network (ANN), the specific initial parameters configured for the nature-inspired meta-heuristic algorithms are detailed below:

- Artificial Neural Network Architecture
- Bat Algorithm (BAT) Parameterization

VIII. ACKNOWLEDGMENT

The authors wish to express their profound gratitude to the Department of Post-Graduation (Computer Science and Information Technology) at MBESs College of Engineering, Ambajogai, for providing the advanced laboratory computing infrastructure, academic resources, and continuous technical support essential to executing this research framework.

Special thanks are extended to the faculty members and fellow research scholars whose insightful critiques and ongoing guidance greatly helped in refining the hybrid optimization architecture. Finally, the authors acknowledge the open-source community and the providers of the standardized SUSAS repository, whose high-fidelity datasets formed the baseline validation for this stress classification system.

REFERENCES

[1] V. V. Yerigeri and L. K. Ragha, "Meta-heuristic approach in neural network for stress detection in Marathi speech, " International Journal of Speech Technology, vol. 22, no. 3, pp. 1-21, 2019.

[2] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," Speech Communication, vol. 40, pp. 5-32, 2003.

[3] F. Eyben, K. Scherer, B. Schuller, et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190-202, 2016.

[4] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in Proc. European Signal Processing Conference (EUSIPCO), pp. 451-455, 2011.

[5] S. H. Xu, J. P. Liu, F. H. Zhang, et al., "A combination of genetic algorithm and particle swarm optimization for vehicle routing problem with time windows," Sensors, vol. 15, no. 9, pp. 21033-21053, 2015.

[6] V. B. Waghmare, R. R. Deshmukh, P. P. Shrishrimal, and G. B. Janvale, "Development of isolated marathi words emotional speech database," International Journal of Computer Applications, vol. 94, no. 4, pp. 19-22, 2014.

[7] Y. Zong, W. Zheng, Z. Cui, and Q. Li, "Double sparse learning model for speech emotion recognition, " Electronics Letters, vol. 52, no. 16, pp. 1410-1412, 2016.