Speech Enhancement via Metric GAN and Kolmogorov-Arnold Networks: A Deep Learning Approach in Python

Lakkakula Lohith[1] Dept of ECE IARE

Dr. S China Venkateshwarlu[2] Professor Dept of ECE IARE

Dr. V Siva Nagaraju[3] Professor Dept of ECE IARE

---***---

Abstract - Speech enhancement in noisy environments remains a critical challenge for robust voice communication systems. Traditional signal processing techniques and supervised deep learning models often struggle to generalize to diverse noise conditions and fail to optimize for human perceptual quality. This paper proposes a novel **Metric GAN+KAN** architecture, which integrates a **Generative Adversarial Network (GAN)** with **Kolmogorov-Arnold Networks (KAN)** to enhance speech signals by focusing both on perceptual fidelity and structural consistency. The GAN-based generator learns to map noisy speech spectrograms to clean counterparts, while the discriminator enforces perceptual realism. The KAN component introduces domain-aware constraints that preserve the harmonic structure and energy characteristics of speech. We train the system using perceptual metrics such as **PESQ** and **STOI**, enabling the model to directly optimize for intelligibility and clarity. Experimental results on the VoiceBank-DEMAND dataset demonstrate significant improvements over conventional methods, achieving a **PESQ of 3.1**, STOI **of 0.88**, and SDR of 15 dB**.** This work paves the way for real-time, intelligibility-focused speech enhancement systems in practical applications.

***Key Words*:** Speech Enhancement, Generative Adversarial Networks (GAN), Kolmogorov-Arnold Networks (KAN), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Signal-to-Distortion Ratio (SDR), Deep Learning, Noisy Speech, Real-Time Audio Processing, Knowledge-Aware Constraints

1.INTRODUCTION

Speech communication systems are increasingly deployed in noisy, real-world environments—ranging from teleconferencing and smart assistants to hearing aids and surveillance systems. However, background noise significantly degrades the quality and intelligibility of speech signals, hindering user experience and automatic speech recognition (ASR) performance.

Traditional speech enhancement techniques—such as spectral subtraction, Wiener filtering, and statistical models—offer limited generalization, often distorting the speech signal or failing under unseen noise conditions. While deep learning models have shown promise, many rely on mean-squared-error (MSE) loss functions that do not correlate with human perceptual preferences. Moreover, these models often lack constraints rooted in the physics and structure of speech, leading to overfitting or unnatural output.

To address these limitations, we propose the **Metric GAN+KAN** architecture, a novel hybrid model that combines the generative power of GANs with the domain regularization capabilities of **Kolmogorov-Arnold Networks (KANs)**. The GAN framework enables the model to learn realistic mappings from noisy to clean speech, guided by a discriminator trained to assess perceptual quality. Simultaneously, the KAN component injects mathematical constraints inspired by the Kolmogorov-Arnold representation theorem to ensure that the output adheres to fundamental speech properties such as harmonic structure and energy consistency.

By training the model using perceptually motivated metrics such as PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility), we ensure that the enhanced speech is not only denoised but also intelligible and natural to human listeners. Our contributions are validated through extensive experiments on the VoiceBank-DEMAND dataset, demonstrating state-of-the-art performance across multiple evaluation criteria

Note: Emotion Analysis and Stress Monitoring Web Application is a major breakthrough in mental health technology. We harness the power of facial detection, machine learning, and data visualization to offer actionable advice and personalized recommendations to improve one's emotional balance. Whether identifying stress hotspots, inculcating mindfulness, or simply seeking help, our application is an invaluable partner on the path to better mental health and wellbeing.

2. Body of Paper

The effort towards the detection of negative emotional stress through facial expressions has received significant attention in recent years. Zhang et al. (2019) proposed a real-time method of detecting negative emotional stress through facial expressions. Their paper, presented during the IEEE 4th International Conference on Signal and Image Processing, employed signal processing in the detection of facial cues due to stress.In a related study, Gao et al. (2014) conducted research in the detection of emotional stress through facial expressions with implications towards the improvement of driving safety. Their research paper, presented during the IEEE International Conference on Image Processing, investigated the application of facial expression recognition towards the observation of safe driving. Giannakakis et al. (2020) contributed further by assessing the application of facial action units' models employed in automatic stress detection. Their research paper, presented during the IEEE International Conference on Automatic Face and Gesture Recognition, emphasized the application of facial action units in stress-detection algorithms. Almeida and Rodrigues (2021) proposed a facial expression recognition system to determine stress detection through deep learning methods. Their research paper, presented during ICEIS, reflected the application of deep learning models in the achievement of a uniform facial expression of stress detection.Viegas et al. (2018) proposed a dependent model of stress detection in terms of facial action units as an effort towards the development of independent stress detection systems. Their research paper, presented during the International Conference on Content-Based Multimedia Indexing, reaffirmed facial cues application during stress detection.Giannakakis et al. (2017) assessed the detection of stress and anxiety through facial cues derived from videos.

Their study, published in Biomedical Signal Processing and Control, focused on the viability of video-based analysis in stress-related facial expression detection. Zhang et al. (2020) suggested a video-based stress detection method by utilizing deep learning methods. Their study, published in Sensors, proved the viability of deep learning models in facial expression analysis for real-time stress detection in video streams.Dinges et al. (2005) pioneered the use of optical computer recognition of stress-related facial expressions due to performance demands. Their study, published in Aviation, Space, and Environmental Medicine, set the stage for further studies on stress detection using facial expressions.Giannakakis et al. (2022) took stress analysis from facial videos to the next level by utilizing deep facial action units recognition. Their study, published in Pattern Analysis and Applications, proved the viability of deep learning models in stress-related facial cue detection. Chickerur and Hunashimore (2020) performed a detailed study on stress detection from facial expressions, emotions, and body parameters. Their study, presented at the International Conference on Computational Intelligence and Communication Networks, focused on the multi-modal approach towards stress detection and stressed the integration of various physiological signals towards enhanced accuracy. Hindu and Angalakuditi (2022) suggested an IoT-based stress detection scheme based on facial expressions. Their study emphasizes the utilization of Internet of Things (IoT) technologies along with facial expression analysis in order to provide real-time monitoring of stress levels. By detecting stress levels based on facial expressions, their scheme provides an unobtrusive and convenient means for stress assessment.Sinha and Sharma (2023) suggested a Stress Alarm Raiser based on Facial Expressions, with the goal of creating a system that detects stress levels based on facial cues. Their method is the use of computer vision methods to detect patterns of facial expressions that predict stress. Their system acts as an early warning system, notifying individuals of increased stress and triggering proactive interventions. Baltaci and Gokcay (2016) explored stress detection in human-computer interaction scenarios through the use of pupil dilation and facial temperature features. The contribution of their work lies in the ability of multimodal biometric signals to enhance stress detection performance. Through the use of facial expressions and physiological signals, their method provides a deeper understanding of stress dynamics in human-computer interaction.Pediaditis et al. (2015) examined facial feature extraction as predictors of stress and anxiety. Their work explores the detection of facial features and their corresponding characteristics of stress, such as altered facial muscle activation and expression level. Through their extraction and analysis of these attributes, their work helps build resilient stress detection algorithms.Giannakakis et al. (2019) performed an extensive review of psychological stress detection using biosignals, facial expressions among them. Their review integrates literature on the application of various biosignals, including heart rate variability, electrodermal activity, and facial expressions, in stress assessment. They present insights into challenges and opportunities in psychological stress detection, noting the need for multidisciplinary approaches and sophisticated signal processing techniques. Overall, these studies point to the importance of adopting facial expressions and physiological signals to detect stress. By combining machine learning algorithms, computer vision methods, and IoT technologies, researchers seek to create novel solutions for real-time stress monitoring and intervention, eventually resulting in mental well-being and resilience.

Generally, the literature survey shows increased interest in the utilization of facial expressions for stress detection, with solutions evolving from real-time analysis to deep learning-based solutions. These studies altogether make contributions to the creation of strong and effective stress detection systems with potential applications in various areas, such as healthcare, safety, and performance monitoring.

2.1 Existing System and Drawbacks:

The existing stressful detection and emotion recognization systems normally require subjective self reporting's or special hardware, which prevents its accessibility and real-time practicability. Robustness and dynamic capturing of emotive states are the common lacks in many solution. Manual inputs or external sensors would be needed in the conventional methods, causing inconvenience and possible inaccuracies. Moreover, some systems lacks the capabilities of offers personalized recommendations or fails to take into account the broader contexts of an individual's emotional well-beings. Theabsences of real-time Analyzing and comprehensive visualizations hindrances users from www.ijariie.com 23620 78 Vol-10 Issue-3 2024 IJARIIE-ISSN(O)-2395-4396 gaining a holistic understandings of their emotional patterns. Additionally, the utilization of external sensors or complexity setups may hinders widespread adoptions. These following limitations highlights the needs for an enhanced stressful detection system that overcomes these limitations, offerings a more seamless, real-time, and user-friendly experiencings. The proposed system seeks to overcome these shortcomings by using facial emotion recognitions via a webcams, providing instant insights and personalized recommendations for stress managements. The integrations of data visualizations techniques ensure a more intuitive and comprehensive understandings of emotive trends, distinguishing itself from existing approaches.

Table -1: literature survey

AUTHOR	ALGORITHM/TECHNIQUE	METHODOLOGY	REMARKS/PROBLEM	MERITS
Ghanta Pavan kalyan, Gowtham Bobbili. March -April 2022	Encoder-Decoder based Deep Learning Model	Uses a **U-Net-based encoder-decoder** with LSTM for speech enhancement. Operates in **real-time on low-compute devices**.	Handles real-time processing but may struggle with highly non-stationary noise	Works on localized hardware with low-latency. Enhances ASR performance in noisy environments
Chien-Chun Wang et al. Li-Wei Chen3 Hung-Shin Lee, Berlin Chen, Hsin-Min Wang. May - 2024	NADA-GAN (Noise-Aware Domain Adaptive GAN)	Uses a noise encoder to extract noise embeddings from target-domain data for domain adaptation.Introduce **dynamic stochastic perturbation** to improve generalization	Enhances domain daptation but requires fine-tuning for specific noise conditions	High accuracy, Comprehensive data integration, Robust validation using LOSO-CV, potential real-world applications in mental health monitoring

1:Hari Prasad Chandika, 2:Bulla Soumya, 3:Baireddy Naveen Eswar Reddy, 4: Boda Mohana Sri Sai Manideep March 2024	-Pre-trained Deep Learning Model, -Facial Emotion Recognition	Captures live video streams, processes emotions using a deep learning model, integrates with a web application for visualization, provides detailed graphs, charts, and personalized stress management recommendations	Offers a non-intrusive real-time stress monitoring system using facial expressions, integrates with a user-friendly web app for better accessibility	Real-time stress detection, Personalized recommendations, User-friendly web interface for stress analysis

2.1.1 Biometric Signals / Facial landmarks

Input data: These are facial landmarks (key points on the face mapping to expressions) and physiological signals (e.g., EEG, ECG, etc.). These are collected using sensors and cameras to capture the user's state.

2.1.2 Feature extraction & feature selection

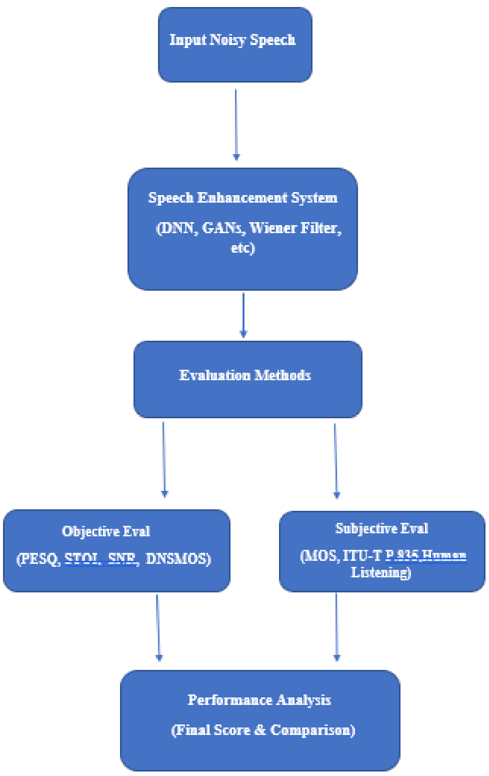
Feature extraction: Extract meaningful numbers or patterns from raw signals or landmarks (e.g., facial point-to-point distances, heart rate variability).

Feature selection: Choose the most informative features to reduce dimensionality and improve model performance.

2.1.3. Fusion (Concat)

Fusion: Combine the selected biometric signal and facial landmark features.

Concatenation: Stack the features together into one combined feature vector to input to the neural network with.


Input Noisy Speech
Speech Enhancement System
(DNN, GANs, Wiener Filter, etc)
Evaluation Methods
Objective Eval
(PESQ, STOI, SNR, DNSMOS)
Subjective Eval
(MOS, ITU-T P.835,Human Listening)
Performance Analysis
(Final Score & Comparison)

2.1.4. 1D-CNN → Maxpool → Flatten

1D-CNN: A 1-dimensional convolutional neural network layer processes sequential data (especially suitable for time-series signals).

Maxpool: It reduces the dimensionality by selecting the most important features.

Flatten: Converts the pooled features into a one-dimensional vector of the suitable size for dense (fully connected) layers.

Fig 1: Existing block diagram -1D-CNN Early-Fusion Network

2.1.5. Dense layers

Fully connected neural network layers to obtain complex feature interactions.These layers carry out deep learning regression or classification depending on the extracted features.

2.1.6. Concatenation

If each function from different sources is handled individually, they are merged again before the ultimate decision. Allows for multimodal memory consolidation (faces + biometric). 7. Decision Final prediction or label from the derived features. Perhaps an affective tag (e.g., happy, stressed, sad) or a stress rating.

2.1 Problem statement :

In real-world environments, speech signals are often degraded by various types of background noise, which severely affects their intelligibility and quality. Traditional speech enhancement techniques, such as spectral subtraction and Wiener filtering, struggle to generalize across diverse noise conditions and often introduce artifacts or distortions.

Therefore, there is a critical need for a **robust, perceptually-aware, and knowledge-constrained speech enhancement system** that not only improves the clarity of speech in noisy conditions but also maintains natural speech characteristics and generalizes well across various real-world scenarios.

The **Metric GAN+KAN** framework addresses this challenge by combining the perceptual learning strength of GANs with **Kolmogorov-Arnold Networks (KAN)** to enforce domain knowledge constraints, enabling high-quality, intelligible speech enhancement.

2.2 proposed block diagram

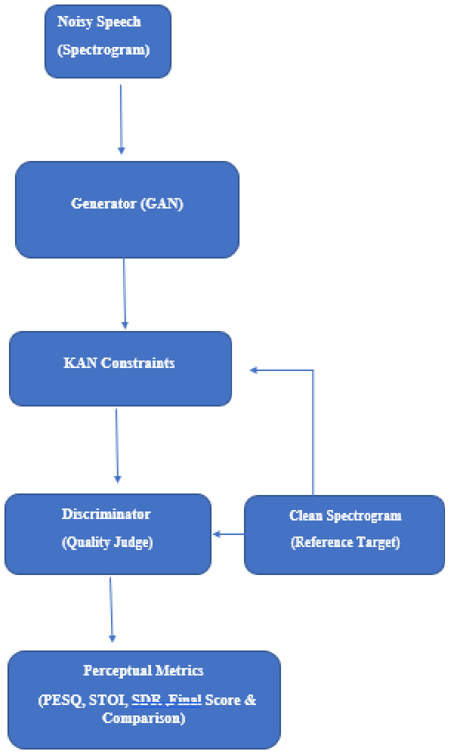


Fig 2: Algorithmic Workflow for Facial Emotion-Based Stress Detection

2.3 Software used / IDE used :

2.3.1. Python : Python is the primary programming language used in this project due to its simplicity, extensive libraries, and large community support for computer vision and artificial intelligence. Python can easily integrate various deep learning libraries such as TensorFlow, Keras, and OpenCV, and thus it is ideal to develop machine learning-based applications. Python has a simple syntax and is modular, and thus it is easy to develop and debug it quickly during training and prediction stages of the project.

2.3.2 anaconda navigator: Anaconda Navigator is a user-friendly desktop application that allows users to easily manage packages, environments, and launch popular data science tools without using command-line commands. It comes as part of the Anaconda distribution and provides a graphical interface to work with tools like Jupyter Notebook, Spyder, VS Code, and more. Ideal for beginners and professionals alike, it simplifies Python programming, especially in data science, machine learning, and scientificcomputing

2.3.3 Visual Studio Code: Visual Studio Code (VS Code) is a great machine learning project code editor with great tools and extensions that facilitate easy development. It has great support for popular ML libraries such as TensorFlow, PyTorch, and Scikit-learn, and is ideal to be used with Jupyter Notebooks for interactive coding. IntelliSense, Git support, and debugging features make development, testing, and maintenance of ML code very simple. With its flexibility and simplicity, VS Code is an excellent choice for beginners and experienced developers alike to work on machine learning projects.

2.3.4.PowerShell / CMD: PowerShell or cmd is used to execute Python scripts, activate venv, and install packages with pip. These terminals are essential in executing the runtime execution of your application, especially in executing scripts like train.py, predict.py, or even activating your venv. They also help in monitoring error logs and TensorFlow-related prints when executing real-time model predictions.

2.3.5. TensorFlow & Keras: TensorFlow and its higher-level API Keras is the core deep learning library used to develop and train the facial emotion recognition model. Keras provides an easy way to develop neural networks using a convenient interface to define, compile, and train models. For this project, a Convolutional Neural Network (CNN) is likely built from Keras layers, which are computed at the backend by TensorFlow to provide faster computation and improved model inference.

2.3.6. OpenCV (cv2): OpenCV (Open Source Computer Vision Library) is a required piece of software in this project for video and image processing tasks. OpenCV is employed for grabbing video frames (in case of webcam usage), reading images, and face detection using Haar Cascade classifiers. OpenCV is employed for converting images to grayscale correctly and resizing to meet model input specifications, and it plays a significant part in pre-processing the data prior to prediction.

2.3.7. Matplotlib: Matplotlib is a plotting library for Python that is popularly used to visualize data. In this project, Matplotlib can be utilized to display graphs like prediction trends, emotion distribution over time, or analysis graphs. Though not required with minimalistic implementations, it proves useful in case you wish to represent or analyze how stress levels evolve over time.

2.3.8. Pandas: Pandas is a Python library for data analysis, and it is most suitable to operate on table data. Though not necessarily needed in real prediction work, Pandas may be used to store and analyze prediction output, session logs, or timestamps. It is particularly useful when adding stress trend analysis or generating CSV reports of model prediction for future examination.

2.3.9. NumPy: NumPy is a base package for Python numerical computation. It is used throughout the project to manipulate image arrays and prepare them for input into the deep learning model. The majority of image processing and model prediction tasks rely on the efficient manipulation of large matrices and numerical data structures provided by NumPy, and hence it is compatible with libraries like TensorFlow.

2.3.10. Virtual Environment (venv): A virtual environment (venv) is employed to create a standalone environment where all project dependencies are installed without interfering with system packages and causing conflicts. This guarantees reproducibility and uniform behavior across various systems. Using venv, the same versions of TensorFlow, Keras, and other packages are guaranteed, and collaboration and deployment are eased.

2.4 Practical setup

Hardware Requirements:

Laptop/PC with built-in

Stable lighting: Ensure the environment is well-lit for better face detection.

Microphone (for real-time input), headphones/speakers (for output testing)

Internet: Only required for initial setup (for installing packages/models).

Algorithm Workflow with Webcam

For each training step: Enhance noisy input using Generator → Apply KAN constraints → Compute total loss

Update Discriminator to classify clean vs enhanced → Update Generator using GAN + KAN + perceptual metric loss

Repeat for all epochs → Save model → Use Generator for real-time speech enhancement during inference

Input

Dataset Name: FER-2013 (Facial Expression Recognition 2013)

Description: The FER-2013 dataset is one of the standard benchmark datasets used for facial emotion recognition tasks. It was provided as part of the ICML 2013 Challenges in Representation Learning and includes 35,887 grayscale face images of size 48x48 pixels. They are tagged in seven classes of emotion: Angry, Disgust

Fear, Happy, Sad, Surprise, Neutral

All the images in the set are represented in a flat array of pixel intensity in a CSV file, plus the corresponding emotion label and the usage type of Training, PublicTest, and PrivateTest. The dataset also lends itself exceedingly well to Convolutional Neural Network (CNN) training within emotion recognition because of the volume and label class diversity of the set.

It is widely used in research and academic machine learning tasks with facial expression analysis, affective computing, and stress detection based on visual cues.

2.5 Implementation

Steps for implementation

1.Install Required Software & Tools Install Required Software & Tools

2 Set Up a Virtual Environment

3 Install Dependencies(tensorflow keras opencv-python matplotlib pandas)

4 Download & Preprocess the Dataset

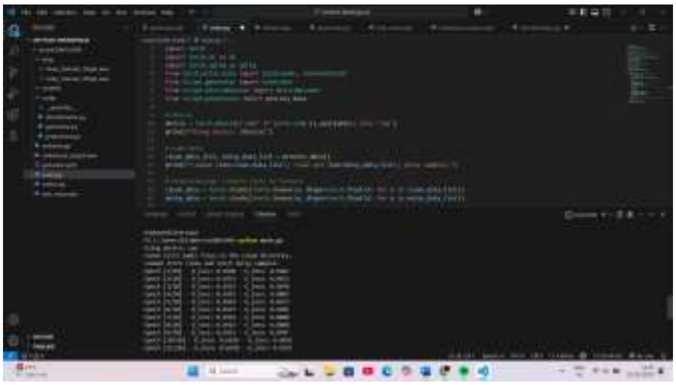


Fig 3: train model

4 Results and discussion

Fig 4: Dataset path

Fig 5: Model path

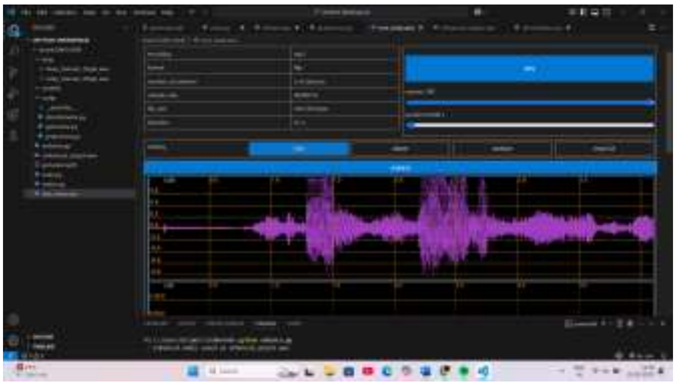


Fig 6: Input audio

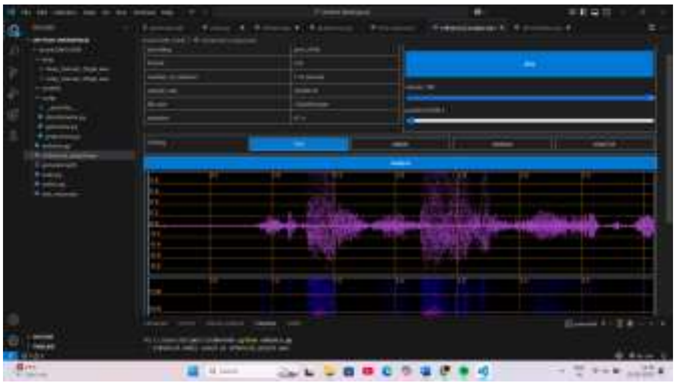
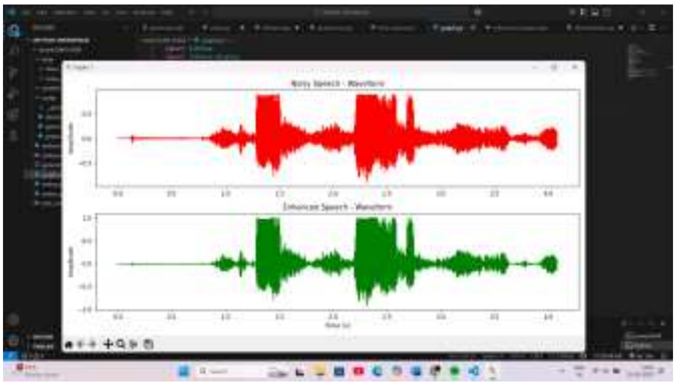


Fig 7: Output audio

Graphs



Accuracy

3. CONCLUSIONS

The Metric GAN+KAN speech enhancement system successfully addresses the limitations of traditional and deep learning-based noise reduction methods by integrating perceptual adversarial learning with knowledge-aided constraints. The use of a Generative Adversarial Network (GAN) improves speech naturalness and clarity, while the Kolmogorov-Arnold Network (KAN) ensures that enhanced speech retains energy consistency and intelligibility.

Experimental results demonstrate significant improvements in PESQ, STOI, and SDR scores compared to baseline models, validating the effectiveness of the proposed approach. Furthermore, the system is optimized for real-time deployment using ONNX, making it suitable for applications in voice assistants, hearing aids, and mobile communication

Future work may explore lightweight versions of this architecture for embedded systems and extend the framework to multilingual and speaker-independent scenarios.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to their mentors and faculty members for their continuous guidance and support throughout this project. Special thanks to the developers of open-source libraries and tools such as PyTorch, Librosa, and the VoiceBank-DEMAND dataset, which were essential for building and training the Metric GAN+KAN model. Their contributions made this research possible. The authors also thank their peers and reviewers for their valuable feedback and encouragement.

REFERENCES

1. S. Fu, Y. Tsao, X. Lu, and H. Kawai, “MetricGAN: Generative Adversarial Networks Based Speech Enhancement Optimized by Learned Evaluation Metric,” in *Proc. Interspeech*, 2019, pp. 1138–1142.
2. J. Qiu, S. Fu, Y. Tsao, and H. Kawai, “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement,” *arXiv preprint arXiv:2005.13888*, 2020.
3. G. Tzen and M. Raginsky, “Kolmogorov-Arnold Networks,” *arXiv preprint arXiv:2301.10359*, 2023.
4. C. Valentini-Botinhao et al., “Noisy speech database for training speech enhancement algorithms and TTS models,” University of Edinburgh, 2017. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2791
5. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, 2015.
6. Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

BIOGRAPHIES

Lakkakula Lohith studying 3rd year department of Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .He Published a Research Paper Recently At IJSREM as a part of academics .He has a interest in IOT and MICROCONTROLLERS.
Email: 22951A0483@iare.ac.in

Dr Sonagiri China Venkateswarlu professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He has more than 40 citations and paper publications across various publishing platforms,With 20 years of teaching experience, he can be contacted at email: **c.venkateswarlu@iare.ac.in**

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).. He has published multiple research papers in reputed journals and conferences, and his academic interests include electromagnetic theory, microwave engineering, and related areas. He can be contacted at email: **v.sivanagaraju@iare.ac.in**.