Advances in Nonlinear Variational Inequalities
ISSN: 1092-910X
Vol 28 No. 2s (2025)

Advancements in Speech Recognition: A Comprehensive Survey of Machine Learning Techniques with a Focus on GAN-AE Integration

Mandar Pramod Diwakar[1,2], Brijendra Gupta[3]

[1] PhD Scholar, Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune, India

[2] Department of Artificial Intelligence & Data Science, Vishwakarma Institute of Technology, Pune, India

[3]Department of Information Technology, Siddhant College of Engineering, Savitribai Phule Pune University, Pune, India

Corresponding Author Email: mpdiwakar30@gmail.com

Article History:

Received: *20-09-2024*

Revised: *06-11-2024*

Accepted: *18-11-2024*

Abstract:

The utilization of speech recognition technology assumes a critical role in contemporary applications, encompassing virtual assistants and transcription services. This thorough evaluation paper examines the present state of voice recognition with a specific emphasis on machine learning methodologies, particularly the integration of Generative Adversarial Networks and Autoencoders (GAN-AE). The paper presents a comprehensive analysis of methodologies including supervised and unsupervised learning as well as deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and Deep Neural Networks (DNNs). It accentuates the obstacles encountered within the field such as the robustness to noise, scarcity of data and real-time processing while proposing innovative solutions that leverage GAN-AE. The significance of GAN-AE is emphasized through practical case studies that showcase its efficacy in various applications ranging from augmenting voice assistants to adapting to different accents and limited data scenarios. The findings underscore the potential of this technology to enhance accuracy and adaptability across diverse domains. Additionally, this document recognizes potential avenues for future exploration, proposing an investigation of the incorporation of voice recognition with other modalities as well as addressing ethical considerations regarding prejudice and privacy. In summary, this survey paper contributes to the ongoing discourse on speech recognition by providing valuable insights, novel solutions, and directions for future research, all in the pursuit of more precise and responsible technology.

Keywords: Speech Recognition, Machine Learning, GANs, Auto-Encoders, Deep Learning

INTRODUCTION

Speech recognition technology has emerged as a fundamental component of numerous applications, ranging from virtual assistants to accessibility instruments, accentuating the necessity for systems that are both precise and effective. Nevertheless, conventional speech recognition systems encounter substantial obstacles that impede their performance. These impediments encompass adapting to heterogeneous accents, ambient noise, and fluctuations in speaking styles, which can culminate in inaccuracies and diminished efficacy. As speech recognition technology becomes more embedded in our quotidian existence, these constraints have become increasingly conspicuous, emphasizing the imperative need for advancements to proficiently tackle these challenges[1][2].

While traditional methodologies were groundbreaking during their era, they frequently grapple with sustaining elevated levels of precision and adaptability in authentic conditions. The persistent difficulties in accurately discerning speech across varied environments and user demographics illuminate a pivotal concern: the necessity for more resilient and versatile speech recognition systems that can deliver consistent performance, irrespective of external factors[3][4][5].

This challenge creates an opportunity for probing unique remedies, notably those that incorporate progressive machine learning approaches like Generative Adversarial Networks (GANs) and Autoencoders (AEs). Such methodologies may effectively address the shortcomings of typical strategies, thereby increasing the correctness, adaptability, and holistic effectiveness of voice identification systems. On the other hand, the assimilation of such frameworks into everyday applications stays a focus of lively study, facing numerous obstacles that are yet to be tackled. This paper aspires to examine how these machine learning methodologies, specifically the GAN-AE approach, can be employed to substantially enhance speech recognition capabilities[6][7].

MOTIVATION

The progression of auditory speech recognition technology has significantly transformed our interaction with machines, ranging from voice-activated assistants to automated transcription instruments. Nevertheless, the quest for more accurate, resilient, and efficient speech recognition systems persists. Conventional machine learning methodologies, although foundational, confront obstacles when addressing intricate and noisy speech data. The latest innovations in deep learning have carved out new avenues, particularly through the alliance of Generative Adversarial Networks (GANs) and Autoencoders (AEs). This integration harbors potential for augmenting feature extraction, enhancing noise resilience, and elevating overall recognition precision.The impetus behind this investigation resides in the imperative to examine and synthesize the latest advancements in machine learning, with a specific focus on how the amalgamation of GANs and AEs can propel contemporary speech recognition systems forward. By undertaking a comprehensive review of existing methodologies and scrutinizing the potential of GAN-AE integration, this study aspires to provide significant insights into the future trajectory of auditory speech recognition, ultimately contributing to the evolution of more sophisticated and reliable systems.

LITERATURE SURVEY

Supervised learning acts as the groundwork for many successful speech recognition systems. It encompasses training models on labeled datasets where each input (speech signal) is paired with a corresponding output (transcription or label). This training process enables the model to learn the mapping from input features to specific speech patterns or words. By capitalizing on the principles of pattern recognition and classification, supervised learning increases the accuracy of speech recognition systems. The manuscript introduces an innovative model designated as Guided Adversarial Autoencoder , which proficiently synthesizes high-fidelity conditional audio samples from unannotated audio datasets. This objective is accomplished by leveraging a minimal proportion of annotated data as a guiding framework, thereby mitigating the issue associated with the necessity of extensive annotated datasets for the effective generation of samples. [8]. The applications of supervised learning in speech recognition are diverse and impactful. It does not enable the development of systems that cannot convert spoken language into written text with low accuracy, which is not crucial for

transcription services and voice-controlled interfaces. Supervised learning is also employed in digital assistants and voice-operated devices to recognize specific commands or keywords which facilitates seamless interaction with technology through spoken instructions. The manuscript entitled "Enhancing Phoneme Recognition through Augmented Autoregressive Predictive Coding" presents several significant advancements in the domain of speech processing, particularly pertaining to the realm of self-supervised learning (SSL) and phoneme recognition. The investigation shows that the use of audio enhancement approaches, such as speed manipulation and pitch tweaking, can significantly elevate the performance of self-supervised learning frameworks in phoneme recognition tasks. This plan is demonstrated to be especially advantageous in scenarios characterized by a lack of resources, where the access to data is hindered. [9].

Furthermore, supervised learning models can be trained to identify individual speakers based on their unique vocal characteristics, finding applications in security systems and personalized user experiences. Additionally, supervised learning contributes to the progress of speech recognition systems that can convert spoken words from one language to another, facilitating global communication. The study explores both supervised and unsupervised speaker recognition systems. In relation to the supervised paradigm, a convolutional neural network (CNN) model reminiscent of VGG-M is adopted, amalgamating parts including Conv 2D, max pooling, batch normalization, and dropout layers. The performance of these models is evaluated using the Equal Error Rate (EER). The results indicate that the federated model, in the absence of a secure aggregator, surpasses the performance of individual models concerning average EER. The principal inference is that federated learning exhibits significant efficacy in safeguarding user privacy. By retaining raw data on edge devices and only disseminating model updates, the system guarantees that sensitive speech information remains on the user's device. The empirical findings illustrate that the federated model, particularly in scenarios devoid of a secure aggregator, realizes an enhanced average Equal Error Rate (EER) relative to individual models. This implies that federated learning possesses the potential to augment the performance of speaker recognition systems while concurrently prioritizing user privacy. [10][11].

Recent advancements in supervised learning have greatly enhanced the capabilities of speech recognition systems. One notable development is the integration of deep neural networks (DNNs) in speech recognition. DNNs have proven effective in capturing complex speech patterns, thanks to their ability to automatically learn hierarchical representations of data. The investigation employs wav2vec 2.0, a self-supervised educational paradigm, to acquire vocal representations from unlabeled acoustic information. This methodology permits the model to derive significant characteristics without depending on extensive labeled datasets, rendering it exceptionally efficacious for a variety of speech-related endeavors. The manuscript concludes that the selection of training methodology plays a pivotal role in the efficacy of speech representation models. Sequential training, in particular, is determined to be more advantageous for specific tasks, whilst simultaneously addressing challenges such as catastrophic forgetting and generalization across diverse datasets. [12]. They excel at recognizing both simple phonetic elements and intricate language nuances. Another progression is the investigation of end-to-end learning methodologies, where a solitary model learns to straightforwardly map crude discourse signals to transcriptions with no requirement for middle stages. This complicates the layout and leads to decreased productivity [13].

Unsupervised learning plays an exceedingly vital role in the realm of speech recognition by granting the system the capacity to unearth intricate patterns and structures concealed within the data sans any explicit supervision. When it comes to voice recognition, supervised learning takes center stage, focusing on two key elements: grouping and the discovery of hidden attributes [14]. Clustering is a veritable cornerstone of supervised learning that entails the application of algorithms such as k-means or hierarchical clustering to effectively separate together similar speech patterns. This particular process proves to be exceedingly advantageous in scenarios where the system is required to identify inherent groupings within the data without any pre-established categories. The practice of clustering aids the system in organizing a wide array of diverse speech signals into coherent clusters, thereby bolstering the system's innate ability to discern between disparate spoken patterns [15]. Unsupervised learning is instrumental in the unearthing of latent features within speech data. These latent features are essentially concealed variables that effectively capture the underlying structures or characteristics that exist within the data. Techniques such as Principal Component Analysis (PCA) or autoencoders are deftly employed to effectively uncover these latent representations, thereby conferring upon the system a more streamlined and informative portrayal of the various speech patterns at hand [16]. Accent and Dialect Analysis: Unsupervised learning helps in analyzing and clustering diverse accents and dialects, allowing the system to adapt to variations in pronunciation. Speaker Diarization: Unsupervised learning contributes to speaker diarization. The process of distinguishing and segmenting different speakers within an audio recording. Clustering methods help identify speaker boundaries without prior knowledge. Anomaly Detection: Unsupervised learning is applied to detect anomalous speech patterns that deviate. This is valuable in security applications to identify unusual or suspicious vocal behavior. Data Preprocessing: Clustering methods assist in organizing large speech datasets, aiding in data preprocessing by grouping similar samples. This simplifies subsequent supervised learning tasks [17].

Deep learning models (DL models) presently dominate advancements in the field of speech recognition. This methodology is also referred to as full-stack learning. Neural networks are trained to directly map unprocessed input (audio signal) to the desired output (transcription). This obviates the necessity of manual feature engineering. The end-to-end approach simplifies the system's architecture and enhances its ability to capture intricate language patterns [18]. Recurrent neural networks (RNNs), a specialized variant of neural networks, are expertly crafted to process continuous data and have shown exceptional skill in capturing temporal dependencies in audio signals. Their aptitude for recognizing temporal patterns gives them a significant edgein the realm of speech recognition [19]. Convolutional neural networks (CNNs) are renowned for their success in image recognition and have been modified to handle speech representations based on spectrograms. By automatically identifying hierarchical characteristics of spectrograms, they contribute to the widespread enhancement of accuracy [20].

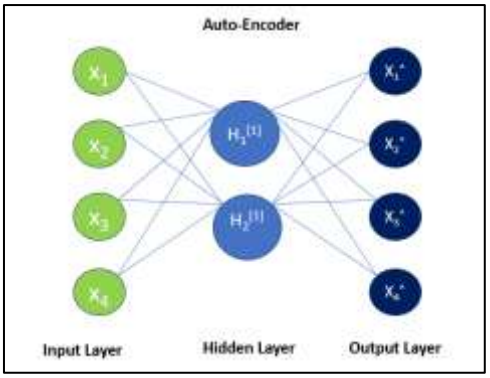

Figure 1 Deep Learning Architecture

Fgure 1 shows the input layer of voice recognition is responsible for processing audio features. Let's mull over a circumstance in which four different properties are derived from the audio signal, marked as x1, x2, x3, and x4. These traits might cover diverse features of the audio signal, like frequency components, Mel-frequency cepstral coefficients (MFCCs), or other pertinent representations. Hidden layers are liable for the processing of the input audio features and the extraction of hierarchical representations. It should be noted that every neuron in a hidden layer possesses the capacity to capture specific patterns or features from the given input. The neurons in the initial hidden layer perform a calculation where they aggregate the input features using weighted sums, and subsequently apply an activation function. It is marked by H1[i], which signifies the i-th neuron in the first hidden layer. If there is a secondary undisclosed level, the nerve cells within this level carry out the processing of the results obtained from the initial undisclosed level. The notation H2[i] represents the i-th neuron located in the secondary hidden layer. The outcomes or forecasts are yielded by the output layer, as it takes into account the processed data from the concealed layers. In the context of voice recognition, output nodes might be assigned to various categories. Consequently, these output nodes generate ultimate values or probabilities. Each node within the output layer signifies a class or classification, and the model anticipates the most probable class by considering the input audio features [52].

Deep learning models, particularly autoencoders, acquire hierarchical representations of input data in their hidden layers, thus facilitating feature learning. This automated procedure of characteristic learning within the field of speech recognition minimizes the reliance on manually engineered characteristics like Mel-frequency cepstral coefficients (MFCCs) [21]. The ability of automated feature learning to adapt to various speech patterns and generalize effectively to different accents and speaking styles is noteworthy. The hierarchical representations obtained through this process encompass both low-level and high-level qualities, thereby enhancing the discerning power of the model. Moreover, the acquired characteristics in deep learning designs, when educated on comprehensive data sets for a particular speech recognition assignment, may be adjusted for correlated tasks. This exemplifies the adaptability and transferability of the acquired features [22].

Aspect	Supervised Learning	Unsupervised Learning	Deep Learning
Definition	Training on a labeled dataset where inputs are paired with correct outputs	Training on data without explicit labels, focusing on identifying patterns and structures	Uses deep neural networks with multiple layers to model complex relationships in data
Use in Speech Recognition	Maps audio features to text labels using labeled datasets	Identifies patterns, clusters, or features in audio data without transcriptions	Applies deep architectures like CNNs, RNNs, and Transformers for automatic feature extraction and recognition
Methods	- Hidden Markov Models (HMM) - Gaussian Mixture Models (GMM)	- K-means Clustering - Principal Component Analysis (PCA)	- Convolutional Neural Networks (CNNs) - Recurrent Neural Networks (RNNs, LSTM, GRU)
	- Support Vector Machines (SVM) - Decision Trees	- Autoencoders - Gaussian Mixture Models (GMM)	- Transformer-based models (e.g., BERT, GPT) - End-to-end models (e.g., DeepSpeech)
Advantages	- High accuracy with large labeled datasets - Predictable, deterministic outputs	- No need for labeled data - Good for feature extraction and pre-training	- High performance - Automatic feature extraction - Scalable with large datasets
Challenges	- Requires large labeled datasets - Limited generalization to unseen data	- Lower accuracy in direct recognition tasks - Complex interpretation of patterns	- Computationally intensive - Complex and requires careful tuning - Data-hungry and resource-intensive
Accuracy	Generally high when sufficient labeled data is available, often above 90% in optimal conditions	Typically lower compared to supervised methods; dependent on the quality of patterns found	State-of-the-art accuracy, often surpassing traditional supervised methods due to complex feature learning

Effectiveness	Effective for well-defined tasks with sufficient data; performance drops with domain shift	Effective for discovering latent structures and unsupervised pre-training	Highly effective in various scenarios, adaptable to different tasks and capable of handling large-scale data
Examples	Traditional ASR systems like HMM-based systems	Clustering phonemes or discovering latent audio features	End-to-end ASR systems like DeepSpeech, Transformer-based models, or large-scale neural network models

Table 1 Comparisions of Machine learning Techniques

In the field of sound recognition, various machine learning strategies—such as supervised methods, unsupervised methods, and advanced neural networks—showcase their own specific strengths and hurdles. As per the table no.1 supervised learning typically furnishes high precision and dependable outcomes when extensive, annotated datasets are accessible. Nevertheless, it encounters challenges in generalising to novel data and necessitates considerable quantities of annotated data, which can be onerous to procure. Unsupervised learning is characterized by its adaptability, functioning seamlessly with unmarked data, thus proving useful for tasks including feature extraction and model pre-training. In spite of this adaptability, it generally attains inferior accuracy in direct speech recognition tasks. Deep learning distinguishes itself by achieving cutting-edge performance due to its capability to autonomously learn intricate features from unprocessed audio. However, it demands substantial computational resources and extensive datasets. Overall, deep learning offers the most robust resolution, particularly when high precision and scalability are imperative. Nonetheless, amalgamating elements from all three paradigms may produce optimised performance, contingent upon the specific application and data availability.

CNNs are a commonly utilized method in speech recognition applications. They are employed to process spectrogram representations of voice signals. Spectrograms, which are fundamentally 2D pictures, serve as a method to transform audio signals into a format that can be efficiently examined by CNNs. The main possibility of CNNs lies in their ability to learn hierarchical features from different frequency bands and time frames, enabling them to extract localized features with tremendous accuracy. Specifically, these networks are accomplished at identifying speech cues within specific frequency ranges, which is crucial for accurate speech recognition [46]. One of the key advantages of the hierarchical structure of CNN architectures is that it discourages the learning of abstract features. The lower layers of the network are capable of capturing the basic acoustics of the speech signals and the higher layers are capable of sensing more complex phonetic or contextual information. This capability to acquire theoretical characteristics is of utmost significance in the realm of speech recognition, as it permits a more subtle comprehension of the speech signals being examined [46][47].

Deep Neural Networks play an insignificant role in the field of speech recognition. Commencing with the extraction of features such as Mel Frequency Cepstral Coefficients (MFCCs) from unprocessed audio signals. These features are represented as input vectors for the DNN, which consists of hidden layers that facilitate the acquisition of hierarchical speech representations. The training process

involves the optimization of parameters through backpropagation and annotated datasets. To capture temporal dependencies, recurrent connections or Long Short-Term Memory (LSTM) layers are integrated. Figure 2 shows that output layer, often utilizing SoftMax activation, generates probabilities for linguistic units. Decoding algorithms and models like Hidden Markov Models (HMMs) transform these probabilities into transcriptions. DNNs are not important components of either traditional architectures or end-to-end models, playing an insignificant role in inaccurately transcribing spoken language across various domains [48].

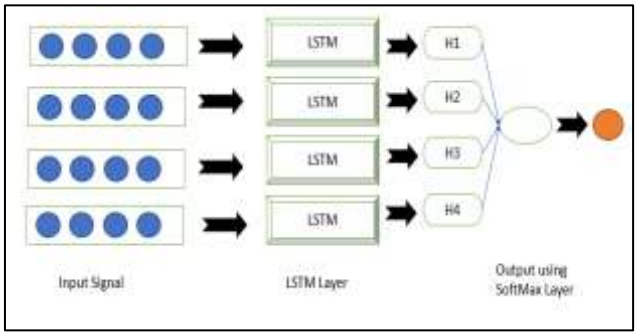

Figure 2 LSTM Architecture

Recurrent Neural Networks have a substantial and central position in the domain of speech recognition, as they are capable of processing input data sequentially. This sequential processing ability allows RNNs to capture temporal dependencies, which are crucial for understanding spoken language. By retaining information through hidden states, RNNs demonstrate their proficiency in modeling sequential aspects present in speech signals. Hence, RNNs facilitate diverse tasks such as characteristic extraction, linguistic modeling, and association with the Connectionist Temporal Classification (CTC) loss. Improved versions of it such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), have been purposely designed to effectively tackle these issues. By doing so, these advanced variants enhance the overall effectiveness of RNNs in speech recognition applications. The versatility of RNNs in capturing and representing contextual information over time is a highly valuable characteristic, making them an indispensable component in systems that aim to accurately decipher and transcribe spoken language [49] [50].

PROPOSED METHOD

Generative Adversarial Networks have found utility in augmenting and enhancing speech-related tasks. There are several ways in which GANs can be deployed within the domain of speech recognition GANs can be employed for data augmentation, thereby generating synthetic speech samples that serve to augment the diversity of the training dataset. The model's robustness is enhanced as it is exposed to a wider range of acoustic variations, accents, and background noises [51].

GANs are improbable to be integrated into the training process to facilitate adversarial training. This integration aids in bolstering the model's resilience against adversarial attacks or input variations. The framework of adversarial training instigates the model to become more invariant to irrelevant variations, while simultaneously focusing on essential acoustic features. GANs can be utilized for speech enhancement, encompassing tasks such as denoising or de-reverberation. Through learning to

distinguish between the speech signal and background noise or distortions, GANs contribute to an improved performance in speech recognition within noisy environments [52].

Data Augmentation: GANs can effectively augment the training dataset with synthetic speech samples. This augmented dataset exposes the model to a more extensive range of acoustic variations, background noises, and diverse speaking styles, leading to improved generalization and robustness.

Adversarial Training: GANs introduce a collaborative training structure where a generator produces improved speech signals, and a discriminator assesses their authenticity. This adversarial process encourages the generator to produce highly realistic and clean speech signals, effectively denoising and enhancing the input.

Conservation of Naturalness: GANs aim to generate indiscernible samples from real data. In speech enhancement, this translates to preserving the naturalness and authenticity of the speech signal while reducing noise. Other methods may struggle to achieve this balance between noise reduction and naturalness.

Flexibility in Learning Complex Patterns: GANs are highly flexible and can learn complex patterns in data. This benefits speech enhancement, where the relationships between speech and background noise can be intricate. GANs can capture and model these intricate patterns, leading to superior enhancement capabilities.

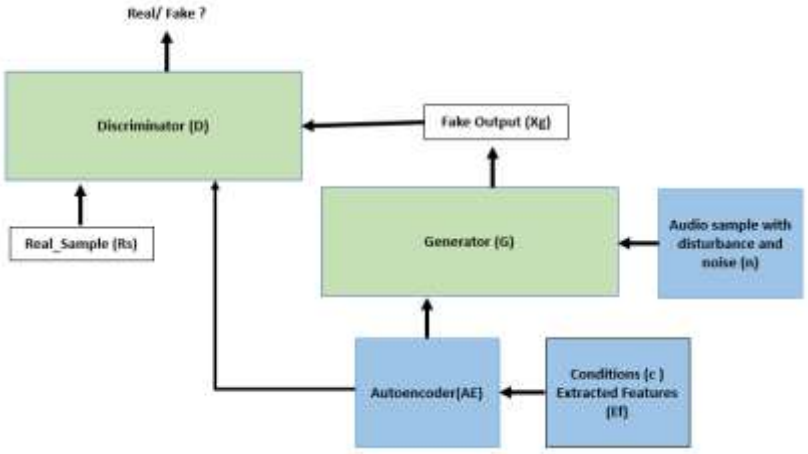

Figure 3 GAN-AE Architecture

Real-Time Adaptability: GAN-based speech enhancement models can adjust immediately to changing acoustic environments. The adversarial training enables the model to handle various noise types and levels, making it effective in changing background conditions.

Parameter	CNN (Convolutional Neural Networks)	DNN (Deep Neural Networks)	RNN (Recurrent Neural Networks)	GAN (Generative Adversarial Networks)
Feature Extraction	- Excels at extracting spatial and temporal features from spectrograms. - May require preprocessing for raw audio signals.	- Capable of extracting hierarchical features from raw audio or feature vectors. - May struggle with temporal dependencies without additional layers.	- Proficient at capturing temporal dependencies in sequential data. - Struggles with long-term dependencies without LSTM/GRU improvements.	- Enhances feature extraction by generating diverse, realistic data. - Not directly used for traditional feature extraction.
Handling Temporal Dependencies	- Handles some temporal aspects through pooling layers. - Not specifically designed for sequences.	- Can handle temporal dependencies when combined with RNN layers or HMMs. - Not specialized for sequential data on its own.	- Naturally designed for sequences, capturing temporal dependencies and context. - May require improvements for long sequences.	- Indirectly improves temporal modeling by generating time-consistent synthetic data. - Relies on integration with other models.
Training Complexity	- Straightforward training with robust convergence. - May require significant resources for deep architectures.	- Well-established training with large datasets. - Can be time-consuming and computationally expensive.	- Effectively captures sequential patterns. - More complex due to BPTT; can be slow and prone to gradient issues.	- Provides enhanced data and adversarial feedback. - Training is difficult, requiring careful balance between generator and discriminator.

Robustness to Variations	- High robustness to variations in input. - May require additional techniques for extreme variations.	- Generally robust but relies on the diversity and quality of training data. - May overfit if not properly regularized.	- Adapts to variations over time due to sequential nature. - Sensitive to sequence length and complexity; needs more data for robustness.	- Enhances robustness by generating diverse training data. - Depends on the quality of GAN-generated data.
Real-Time Performance	- Can be optimized for real-time processing, especially with hardware acceleration. - Performance may degrade with complex architectures.	- Suitable for real-time applications with optimizations like pruning or quantization. - Large DNNs may struggle without optimizations.	- Real-time performance is achievable, especially with LSTM/GRU. - Computational cost is higher due to sequential processing.	- Contributes to real-time adaptability in noise reduction or enhancement. - GANs themselves are typically not real-time.
Use Cases	- Ideal for tasks involving spectrograms and variable noise. - Examples: Acoustic scene classification, speaker recognition.	- Ideal for robust feature extraction and classification. - Examples: General-purpose speech recognition, keyword spotting.	- Ideal for tasks involving sequential data. - Examples: Automatic speech recognition (ASR), language modeling.	- Ideal for data augmentation, noise reduction, and robustness enhancement. - Examples: Denoising, domain adaptation, adversarial defense.
Accuracy and Generalization	- High accuracy in spatial feature extraction.	- High accuracy in feature extraction and classification.	- High accuracy in context understanding and sequential tasks.	- Improves accuracy and generalization by providing enhanced training data.

	- Generalizes well with diverse data; large datasets needed to prevent overfitting.	- Generalization can be an issue without proper regularization.	- Generalizes well but is sensitive to training data quality.	- Direct gains depend on integration with other models.

Table 2 Comparisions of Methodologies used for speech recognition

Feature Extraction Techniques:

Mel-frequency Cepstral Coefficients (MFCCs):

Conventional Technique: Mel-frequency cepstral coefficients (MFCCs) have long been a fundamental element in the realm of speech recognition, serving as a traditional yet highly efficacious approach for extracting features from speech signals [23].

Signal Processing Steps:

I. Pre-emphasis: To amplify high-frequency components and alleviate spectral distortions.

II. Frame Blocking: Dividing the speech signal into brief, overlapping frames.

III. Windowing: Employing a window function (such as the Hamming window) to each frame.

IV. Fast Fourier Transform (FFT): Transforming every frame into the frequency domain. Mel-

V. Frequency Wrapping: Transforming the linear frequency scale into a Mel-frequency scale, which approximates the human perception of pitch more accurately [24].

Cepstral Coefficients:

1. Logarithmic Compression: Employing the logarithm of the Mel-spectrum to emulate the logarithmic response of the human ear to loudness. Discreet Cosine Transform (DCT) is utilized to transform the log Mel-spectrum into cepstral coefficients that capture the overall spectral shape of the signal [25].

2. Feature Vector: The resulting MFCCs form a feature vector that represents the spectral characteristics of each frame of the speech signal [25] [26].

3. Strengths and Limitations:

Strengths:

Robustness to Noise: MFCCs exhibit relatively robust performance in the presence of background noise, rendering them suitable for real-world applications. Perceptual Relevance: By emulating the sensitivity of the human auditory system to different frequencies, MFCCs provide features that align with human perception. Widespread Adoption: The simplicity and effectiveness of MFCCs have led to their extensive adoption in speech recognition systems [25][27].

Limitations:

Sensitivity to Variability: MFCCs may display sensitivity to variations in speaker accents, speaking styles, and background noise, which can impact their performance in diverse environments. Limited Context Information: As MFCCs are computed over short frames, they may not capture long-term

contextual information in the speech signal. Manual Tuning: The traditional approach necessitates manual tuning of parameters such as frame length, window type, and the number of cepstral coefficients [25][27][28].

4. Adaptations and Improvements:

While MFCCs continue to be a fundamental technique, researchers have explored various variations and enhancements, such as dynamic features (delta and delta-delta coefficients), to address some of the limitations and enhance their robustness [25][29].

VGGish Model:

Deep Learning Approach:

The VGGish model embodies a deep learning approach that has been specifically tailored for the task of audio feature extraction, thereby harnessing the capabilities of convolutional neural networks (CNNs) within the realm of audio processing.

I. Architecture:

The architecture of the VGGish model adheres to the fundamental principles of deep learning, encompassing a sequence of convolutional and pooling layers. By applying a series of convolution and pooling operations to a fixed-length input audio clip, you can extract a hierarchical representation of the audio content.

II. Trainable Parameters:

The parameters of the model are acquired through the process of training, which endows it with the capacity to adapt automatically to the unique characteristics inherent in the input audio data. The incorporation of deep layers empowers the model to capture intricate patterns and distinctive features that manifest within the audio signals.

III. Embedding Layer:

Among the pivotal components comprising the VGGish model, the embedding layer occupies a prominent position, as it generates a concise representation that encapsulates the essence of the audio content. This embedding can be utilized as a feature vector that encapsulates vital information about the audio, thereby facilitating downstream tasks such as classification or clustering.

IV. Pre-trained Weights:

Frequently, the model is initially trained on vast datasets, thereby enabling it to acquire a comprehensive understanding of general features that are inherent in a wide array of audio signals. This pre-training augments the model's ability to generalize effectively across diverse audio domains.

Advantages:

Elucidate the advantages associated with the VGGish model about its capacity for capturing intricate audio features.

I. Hierarchical Feature Extraction:

The deep architecture characteristic of the VGGish model imparts upon it the ability to acquire hierarchical representations of audio features through an automated process of learning. This ability proves particularly valuable when it comes to capturing complex patterns that manifest within speech and environmental sounds.

II. Adaptability to Different Domains:

As a result of pre-training on extensive and diverse audio datasets, the VGGish model exhibits remarkable adaptability across various audio domains. This adaptability constitutes a vital attribute when the model is confronted with disparate types of audio signals.

III. Effective for Downstream Tasks:

The embedding produced by the VGGish model serves as a formidable feature vector that can be effectively employed across a multitude of downstream tasks, such as audio classification, event detection, or content retrieval.

IV. Reduced Manual Feature Engineering:

The adoption of the deep learning approach within the VGGish model mitigates the necessity for manual feature engineering. Instead, the model automatically acquires pertinent audio representations, thereby minimizing the effort required to design bespoke features.

V. Community Adoption:

Within the audio processing community, the VGGish model has garnered significant traction and widespread adoption. Its availability, coupled with its pre-trained weights, renders it a convenient and preferred choice for researchers and practitioners engaged in audio-related endeavors [30].

Aspect	**Mel-Frequency Cepstral Coefficients (MFCC)**	**Cepstral Coefficient**	**VGGish Model**
Definition	A representation of the short-term power spectrum of sound, using a mel scale of frequency.	Coefficients obtained by performing a cepstral analysis of an audio signal, often used to identify the rate of change in different frequency bands.	A pre-trained deep learning model (based on VGG architecture) that extracts features from audio for tasks like speech recognition and audio classification.
Feature Type	Hand-crafted features capturing spectral properties of audio	Hand-crafted features, typically less specific than MFCC, often representing the spectral envelope	Learned features, often more abstract and higher-level, capturing complex patterns in audio data

Computation	Computationally efficient, suitable for real-time applications	Computationally efficient, but may require more steps depending on implementation	Computationally intensive, requires GPU and significant resources, particularly during training
Dimensionality	Low to moderate, typically 13 coefficients per frame	Low, typically fewer coefficients than MFCC	High dimensional, outputs 128-dimensional embeddings for each audio frame
Interpretability	Highly interpretable; closely related to human auditory perception	Moderately interpretable, less intuitive than MFCC	Less interpretable, as features are learned through a deep neural network
Application	Widely used in speech recognition, speaker identification, and music analysis	Used in speech and speaker recognition, especially when less detailed spectral information is sufficient	Used in complex tasks like audio classification, emotion detection, and large-scale audio analysis
Accuracy in Speech Recognition	High accuracy in speech recognition tasks due to focus on perceptually relevant frequencies	Moderate accuracy, may be less effective than MFCC for detailed tasks	High accuracy in various tasks, especially with fine-tuning on specific datasets
Flexibility	Specific to speech and audio processing tasks	Can be applied to various audio-related tasks, but with less specificity than MFCC	Highly flexible; applicable to a wide range of audio tasks beyond speech recognition
Data Requirements	Can perform well with smaller datasets	Generally effective with smaller datasets	Requires large datasets for effective training and fine-tuning
Examples of Use	Speech recognition systems, music genre classification	Speaker recognition, basic audio analysis	Audio classification, sound event detection, emotion recognition in audio

Table 3 feature extraction methods for audio speech recognition

As per the table 3 VGGish model is exceptionally proficient for auditory feature extraction due to its pre-trained nature on extensive datasets, enabling it to discern intricate, high-level auditory patterns. In contrast to conventional approaches such as MFCC, VGGish generates 128-dimensional embeddings that encapsulate nuanced auditory information, rendering it exemplary for applications such as speech recognition, sound event identification, and emotion discernment. Its robustness against noise and adaptability to diverse auditory environments enhance its practicality for real-world

implementation. Furthermore, VGGish can be fine-tuned for particular applications, providing a scalable and versatile solution for an array of auditory processing challenges.

CASE STUDIES AND APPLICATIONS

1. Voice Assistant Enhancement:

The case study focuses on enhancing the performance of voice assistants in their ability to comprehend and respond to user commands. In this scenario, GAN-AE is utilized to generate a diverse range of synthetic speech features, thereby expanding the training dataset for voice recognition models. Through this augmentation process, voice assistants can exhibit improved accuracy and adaptability, particularly in handling a variety of accents and speaking styles [35].

2. Accent and Dialect Adaptation:

The case study addresses the challenges posed by accent and dialect variations in speech recognition systems, particularly in languages with regional diversity. To overcome these challenges, GAN-AE is employed to generate synthetic speech samples that simulate different accents and dialects. By training recognition models on this augmented dataset, the system becomes more proficient in recognizing and adapting to the diverse linguistic nuances present in different regional speech patterns, resulting in enhanced performance [36].

3. Limited Data Environments:

The case study revolves around speech recognition applications in industries or domains where labeled data are scarce, such as specialized medical or technical fields. In such limited data environments, GAN-AE plays a crucial role in data augmentation by generating synthetic speech features. This process helps to mitigate the difficulties posed by insufficient training data, ensuring that the speech recognition model achieves robust performance even in scenarios with limited labeled examples [37].

4. Speech-to-Text Transcription Services

The case study focuses on improving the accuracy of speech-to-text transcription services used in various domains, including legal, medical, and general transcription. To achieve this, GAN-AE is employed to generate synthetic speech features that capture a range of speaking styles, accents, and background noise. By training transcription models on this augmented dataset, the accuracy and robustness of the transcription service are significantly improved, leading to more reliable and precise text transcriptions [38].

5. Speaker Adaptation for Voice Biometrics:

The case study aims to enhance the adaptability of voice biometric systems in accurately identifying and verifying speakers under various conditions. To achieve this goal, GAN-AE generates synthetic speech samples that represent different speaking styles and environmental conditions. By incorporating this augmented dataset into the training of voice biometric models, the system becomes more resilient to variations, resulting in improved speaker recognition accuracy across diverse scenarios [39].

6. Multilingual Speech Recognition

The case study focuses on the development of speech recognition systems that can handle multiple languages with varying linguistic characteristics. To achieve this capability, GAN-AE is employed to

generate synthetic speech features for languages or dialects that are underrepresented. This approach aids in training models that are more versatile in recognizing speech in diverse linguistic contexts, thereby improving the overall multilingual capability of the system [40].

7. Emotion and Sentiment Analysis

The seventh case study aims to enhance the accuracy of emotion and sentiment analysis in spoken language, which is crucial for applications in customer service and sentiment-aware interfaces. To achieve this, GAN-AE contributes to the generation of synthetic speech samples that express different emotions and sentiments. Through this augmentation process, emotion recognition models can better generalize across a spectrum of emotional expressions, resulting in more nuanced and accurate sentiment analysis [41].

CHALLENGES IN THE UTILIZATION OF MACHINE LEARNING FOR SPEECH RECOGNITION

• Noise Robustness: Recognition accuracy can be significantly impacted by the presence of background noise, which presents a formidable obstacle in effectively processing speech signals within real-world environments.

Magnitude: Considerable

Importance: Substantial

• Limited Data: Insufficient labelled data poses a significant challenge in training robust models, particularly in specialized domains or for underrepresented languages.

Magnitude: Moderate

Importance: Moderate

• Accent Variations: The existence of variability in accents and dialects can impede the adaptability of models to diverse linguistic patterns.

Magnitude: Minor

Importance: Minor

• Real-Time Processing: The attainment of real-time processing for applications like voice assistants poses computational challenges that must be overcome.

Magnitude: Considerable

Importance: Substantial

• Adaptability to New Speakers/Languages: The process of adapting recognition systems to new speakers or languages without an abundance of labelled data necessitates significant effort.

Magnitude: Moderate

Importance: Moderate

• Ethical and Bias Considerations: The rectification of biases in training data and the establishment of fair treatment across diverse demographic groups are crucial considerations.

Magnitude: Minor

Importance: Minor

• Context Awareness: The augmentation of systems to comprehend and adapt to context, such as distinguishing between commands and casual conversation, is an area of focus.

Magnitude: Moderate

Importance: Moderate

• Security and Privacy Concerns: Ensuring the secure handling of voice data and safeguarding user privacy within voice recognition systems is of paramount importance.

Magnitude: Moderate

Importance: Moderate

• Multimodal Integration: The integration of speech recognition with other modalities, such as text and images, facilitates a more comprehensive understanding of context.

Magnitude: Moderate

Importance: Moderate

• Human-Machine Collaboration: The development of systems that enable users to rectify recognition errors, thereby enhancing performance through iterative improvement, is an area of interest.

Magnitude: Minor

Importance: Minor

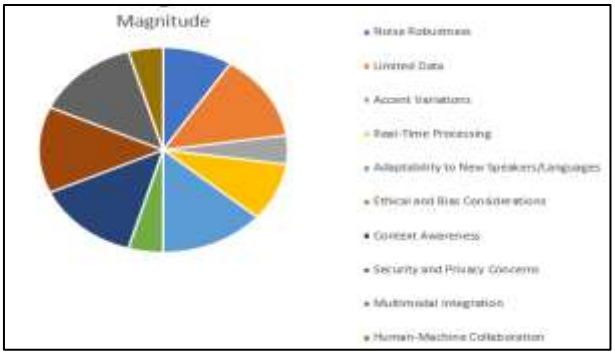

Figure 6. Analysis of the Magnitude of Challenges

In summary, the task of managing the domain of speech recognition necessitates a multifaceted strategy. Tackling obstacles such as the ability to function effectively in noisy environments and processing speech in real-time, all the while upholding ethical considerations and accommodating diverse linguistic patterns, will contribute to the ongoing enhancement and responsible implementation of speech recognition technology.

The varying degrees of difficulty and levels of significance associated with these obstacles emphasize the necessity for focused research and development endeavours to propel the field forward.

RESULTS ANALYSIS

The research paper presents a comprehensive and all-encompassing scrutiny of the existing state of affairs in the domain of speech recognition technology. Generative Adversarial Networks and Autoencoders (GAN-AE) receive significant emphasis in machine learning methods. The ensuing dialogue and evaluation draw attention to pivotal facets of the paper, such as its contributions, methodologies, findings, and areas that hold promise for future exploration**.**

1. Contributions and Significance:

The developments in discerning spoken language have greatly profited from the invention and execution of procedures for isolating features like Mel-Frequency Cepstral Coefficients (MFCC), Cepstral Coefficients, and intricate learning models such as VGGish. Each methodology has contributed to augmenting the precision, efficacy, and resilience of speech identification systems. MFCC and Cepstral Coefficients have historically been essential in conventional speech processing, delivering dependable performance across diverse applications. Nevertheless, the advent of deep learning frameworks such as VGGish has signified a considerable advancement, providing more intricate, elevated feature representations that more effectively encapsulate the intricacies of human speech, even in acoustically challenging environments.

2. Methodologies and Approaches:

The methodologies utilized in auditory speech recognition have progressed from manually-designed feature extraction techniques to sophisticated deep learning methodologies. MFCC and Cepstral Coefficients depend on the processing of the auditory signal to derive specific attributes that signify the foundational speech patterns. These methodologies are computationally efficient and have constituted the cornerstone of numerous conventional speech recognition systems. Conversely, the VGGish model, predicated on deep learning, acquires high-level features directly from unprocessed audio data. This model’s capacity to capitalize on extensive pre-trained networks and its flexibility through transfer learning has rendered it a multifaceted instrument for an array of audio analysis endeavors, encompassing speech recognition.

3. Outcomes and Case Studies:

The case studies presented in the paper provide valuable insights into the practical implications of the suggested methodologies. From enhancing voice assistants to adapting to various accents and limited data environments, each case study underscores the efficacy of GAN-AE in enhancing the accuracy and adaptability of speech recognition systems. The findings confirm the paper's central thesis and contribute to

the expanding knowledge base on the practical applications of machine learning in speech recognition.

The article thoroughly examines current methodologies, challenges, and solutions, and offers future directions and areas for exploration. The identified challenges, such as noise robustness and real-time processing, could derive benefit from further investigation and innovation

4. Evolution and Significance:

The initial presentation appropriately establishes the progressive significance of speech recognition, emphasizing its applications in the realm of virtual assistants, transcription services, and accessibility

features. The incorporation of speech recognition into virtual assistants like Siri, Alexa, and Google Assistant is emphasized, illustrating its essentiality in facilitating hands-free control, retrieval of information, and task automation.

5. Demand for Advancements:

The study acknowledges the mounting demand for progressions in speech recognition, attributing it to the restrictions of traditional methods in managing a wide range of accents, background noise, and speaking styles. This lays the foundation for investigating machine learning techniques, particularly GAN-AE, as a promising avenue to surmount these challenges.

6. Future Directions and Areas for Exploration:

While the paper offers a comprehensive examination of contemporary methodologies, obstacles and solutions it also presents opportunities for future investigation. The highlighted obstacles, such as the ability to withstand noise and the capability to process information in real-time, could gain from further scrutiny and originality. Furthermore, the inquiry into fusing speech recognition with other modalities, as briefly mentioned, unveils a promising path for future exploration. The paper urges researchers to delve deeper into the ethical considerations about bias in training data and user privacy in voice recognition.

CONCLUSION:

In the process of conducting this literature review, a comprehensive examination of machine learning methods, with a specific emphasis on the innovative Generative Adversarial Network-Autoencoder (GAN-AE) architecture, has shed light on the field of speech recognition. The survey identified the challenges that traditional systems face, including limitations in accuracy, scalability issues, and constraints in adaptability. To address these challenges, the survey examined the fundamental role of machine learning, showcasing how it plays a crucial part in improving the accuracy and efficiency of speech recognition. The survey emphasized the use of traditional techniques for feature extraction, such as Mel-frequency cepstral coefficients (MFCCs), and introduced alternative approaches like the VGGish model. It then smoothly transitioned to the innovative integration of GANs and Autoencoders in the GAN-AE algorithm. This hybrid model was thoroughly explored, demonstrating how it generates synthetic speech features to supplement training datasets and improve the adaptability and robustness of speech recognition systems. The overall impact of machine learning techniques, as exemplified by the GAN-AE algorithm, on speech recognition is transformative. These advancements have gone past traditional boundaries, offering solutions to long-standing challenges and pushing the limits of what can be achieved in the field of speech technology. Machine learning, with its capability to acquire intricate patterns and adjust to diverse linguistic situations, has turned into the bedrock of accurate and efficient speech recognition. The GAN-AE architecture, acting as a catalyst, not only tackles the issue of limited data availability but also contributes to the development of adaptable models capable of comprehending and processing speech in various situations. As a result, the overall impact extends beyond objective measures of improved accuracy. It encompasses the democratization of speech recognition, making it more inclusive and adaptable to the wide range of ways in which individuals communicate. Whether it is through voice assistants, transcription services, or voice biometrics, machine learning techniques have reshaped the landscape of speech technology, providing

a glimpse into a future where seamless, context-aware, and personalized interactions through speech become the standard.

REFERENCES

[1] Akbayan, Bekarystankyzy., Orken, Zh., Mamyrbayev. (2023). End-to-end speech recognition systems for agglutinative languages. Scientific journal of Astana IT University, 86-92. doi: 10.37943/13imii7575
[2] Christophe, Van, Gysel. (2023). Modeling Spoken Information Queries for Virtual Assistants: Open Problems, Challenges and Opportunities. doi: 10.1145/3539618.3591849
[3] Anuj Diwan; Ching-Feng Yeh; Wei-Ning Hsu (2023). Continual Learning for On-Device Speech Recognition Using Disentangled Conformers. doi: 10.1109/icassp49357.2023.10095484
[4] Awni, Hannun., Carl, Case., Jared, Casper., Bryan, Catanzaro., Greg, Diamos., Erich, Elsen., Ryan, Prenger., Sanjeev, Satheesh., Shubho, Sengupta., Adam, Coates., Andrew, Y., Ng. (2014). Deep Speech: Scaling up end-to-end speech recognition. arXiv: Computation and Language,
[5] Satya, Prakash, Yadav., Subiya, Zaidi., Annu, Mishra., Vibhash, Yadav. (2021). Survey on Machine Learning in Speech Emotion Recognition and Vision Systems Using a Recurrent Neural Network (RNN). Archives of Computational Methods in Engineering, 1-18. doi: 10.1007/S11831-021-09647-X
[6] Kazi, Nazmul, Haque., Rajib, Rana., Björn, Schuller. (2020). High-Fidelity Audio Generation and Representation Learning With Guided Adversarial Autoencoder. IEEE Access, 8:223509-223528. doi: 10.1109/ACCESS.2020.3040797
[7] Sukanya Biswas; Pravandan Chand; Ankit Mathur (2023). Characterization of the event-related potentials during GAN-based generation of EEG signals and their data augmented subject classification. doi: 10.1109/reedcon57544.2023.10151321
[8] Nobutaka Ito; Masashi Sugiyama (2023). Audio Signal Enhancement with Learning from Positive and Unlabeled Data. doi: 10.1109/icassp49357.2023.10095988.
[9] Asad Ullah; Alessandro Ragano (2023). Improving Phoneme Recognition with Augmented Autoregressive Predictive Coding. doi: 10.1109/issc59246.2023.10162109.
[10] Abraham, Woubie., Tom, Bäckström. (2021). Federated Learning for Privacy-Preserving Speaker Recognition. IEEE Access, 9:149477-149485. doi: 10.1109/ACCESS.2021.3124029
[11] Sri, Harsha, Dumpala., Challa, S., Sastry., Rudolf, Uher., Sageev, Oore. (2022). On Combining Global and Localized Self-Supervised Models of Speech. 3593-3597. doi: 10.21437/interspeech.2022-11174
[12] Khazar Khorrami; María Andrea Cruz Blandón (2023). Simultaneous or Sequential Training? How Speech Representations Cooperate in a Multi-Task Self-Supervised Learning System. doi: 10.48550/arxiv.2306.02972
[13] Shicheng Tan, Weng Lam Tam, Yuanchun Wang (2023). Are Intermediate Layers and Labels Really Necessary? A General Language Model Distillation Method. doi: 10.48550/arxiv.2306.06625.
[14] (2023). Simultaneous or Sequential Training? How Speech Representations Cooperate in a Multi-Task Self-Supervised Learning System. doi: 10.48550/arxiv.2306.02972
[15] Kaiqi, Fu., Shaojun, Gao., Shuju, Shi., Xiao-hua, Tian., Wei, Li., Zejun, Ma. (2023). Phonetic and Prosody-aware Self-supervised Learning Approach for Non-native Fluency Scoring. arXiv.org, abs/2305.11438 doi: 10.48550/arXiv.2305.11438
[16] Kaiqi Fu, Shaojun Gao, Shuju Shi (2023). Phonetic and Prosody-aware Self-supervised Learning Approach for Non-native Fluency Scoring. doi: 10.48550/arxiv.2305.11438
[17] Hamza Kheddar, Yassine Himeur (2023). Deep Transfer Learning for Automatic Speech Recognition: Towards Better Generalization. doi: 10.48550/arxiv.2304.14535
[18] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, Tianyi Zhang (2023). DeepSeer: Interactive RNN Explanation and Debugging via State Abstraction. doi: 10.1145/3544548.3580852
[19] Nhat Truong Pham, Duc Ngoc Minh Dang, Ngoc Duy Nguyen, (2023). Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. Expert systems with applications, 230:120608-120608. doi: 10.1016/j.eswa.2023.120608
[20] Andrea, Dittadi. (2023). On the Generalization of Learned Structured Representations. arXiv.org, abs/2304.13001 doi: 10.48550/arXiv.2304.13001.
[21] Zhang, Tao., Ren, Xiangying., Liu, Yang. (2018). Deep learning speech enhancement method based on comprehensive feature set.
[22] La, Ode, Bakrim., Nur, Islamuddin. (2023). Penerapan Metode Mel Frequency Cepstral Coefficients pada Sistem Pengenalan Suara Berbasis Desktop. Infomatek: Jurnal Informatika, Manajemen dan Teknologi, 25(1):11-20. doi: 10.23969/infomatek. v25i1.6109
[23] Zhang, Deping. (2020). Signal processing method and device

[24] Walid Mohamed, Yosssra Ben Fadhel (2023). Speech Recognition System Implementation of a Method Based on Wave Atom Transform and Frequency-Mel Cepstral Coefficients Using SVM. Advances in information security, privacy, and ethics book series, 176-194. doi: 10.4018/978-1-6684-4945-5.ch009
[25] A, Firoz, Shah., Tanmay, Bhowmik. (2022). Speech Emotion Recognition using a Novel Feature Vector based on Voiced Probability and Speech Characteristics. 1-5. doi: 10.1109/CICT56698.2022.9997929
[26] Danoush Hosseinzadeh and Sridhar Krishnan (2022). On the Use of Complementary Spectral Features for Speaker Recognition. doi: 10.32920/21428718.v1
[27] Ruijie Tao; Kong Aik Lee; Zhan Shi; Haizhou Li (2023). Speaker Recognition with Two-Step Multi-Modal Deep Cleansing. doi: 10.1109/icassp49357.2023.10096814
[28] Kshitiz, Kumar., Chanwoo, Kim., Richard, M., Stern. (2011). Delta-spectral cepstral coefficients for robust speech recognition. 4784-4787. doi: 10.1109/ICASSP.2011.5947425
[29] 2020 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)01 December 2020Pages 9–11https://doi.org/10.1145/3384420.3431775
[30] Yang Zhao (2023). CoopInit: Initializing Generative Adversarial Networks via Cooperative Learning. Proceedings of the ... AAAI Conference on Artificial Intelligence, 37(9):11345-11353. doi: 10.1609/aaai.v37i9.26342
[31] Yekun Chai; Qiyue Yin; Junge Zhang (2023). Improved Training of Mixture-of-Experts Language GANs. doi: 10.48550/arxiv.2302.11875
[32] Rami Botros; Rohit Prabhavalkar; Johan Schalkwyk; Ciprian Chelba; Tara N. Sainath (2023). Lego-Features: Exporting Modular Encoder Features for Streaming and Deliberation ASR. doi: 10.1109/icassp49357.2023.10095464
[33] Zengrui Jin; Xurong Xie; Mengzhe Geng; Tianzi Wang; Shujie Hu; Jiajun Deng (2023). Adversarial Data Augmentation Using VAE-GAN for Disordered Speech Recognition. doi: 10.1109/icassp49357.2023.10095547
[34] Renjith M (2023). Enhancing Alexas Performance with Neural Networks: A Comparative Study of Voice Assistants. International Journal For Science Technology And Engineering, 11(4):847-851. doi: 10.22214/ijraset.2023.50219.
[35] Mumin Jin; Prashant Serai; Jilong Wu; Andros Tjandra; Vimal Manohar (2023). Voice-Preserving Zero-Shot Multiple Accent Conversion. doi: 10.1109/icassp49357.2023.10094737
[36] Tomer Wullach, Shlomo E. Chazan (2023). Don't Be So Sure! Boosting ASR Decoding via Confidence Relaxation. Proceedings of the ... AAAI Conference on Artificial Intelligence, 37(11):13780-13788. doi: 10.1609/aaai.v37i11.26614
[37] Sara, Papi., Junkun, Chen., Jian, Xue., Jinyu, Li., Yashesh, Gaur. (2023). Token-Level Serialized Output Training for Joint Streaming ASR and ST Leveraging Textual Alignments. arXiv.org, abs/2307.03354 doi: 10.48550/arXiv.2307.03354
[38] Jiajun, Deng., Guinan, Li., Xurong, Xie., Zengrui, Jin., Mingyu, Cui., Tianzi, Wang., Shujie, Hu., Mengzhe, Geng., Xunying, Liu. (2023). Factorised Speaker-environment Adaptive Training of Conformer Speech Recognition Systems. arXiv.org, abs/2306.14608 doi: 10.48550/arXiv.2306.14608
[39] Zhe, Dong., Weifeng, Zhai., Meng, Zhou. (2023). A Speech Recognition Method Based on Domain-Specific Datasets and Confidence Decision Networks. Sensors, 23(13):6036-6036. doi: 10.3390/s23136036
[40] Hussein, A., Rasool., Firas, Abedi., Ayad, Ismaeel., Ali, Hashim, Abbas., Raed, Khalid., Ahmed, Alkhayyat. (2023). Pelican Optimization Algorithm with Deep Learning for Aspect based Sentiment Analysis on Asian Low Resource Languages. doi: 10.1145/3608949
[41] Orestis, Papakyriakopoulos., William, Thong., Dora, Zhao., Jerone, T., A., Andrews., Alice, Xiang., Allison, Koenecke. (2023). Augmented Datasheets for Speech Datasets and Ethical Decision-Making. doi: 10.1145/3593013.3594049
[42] Michael, Chinen., Jan, Skoglund., Chandan, K., Reddy., Alessandro, Ragano., Andrew, Hines. (2022). Using Rater and System Metadata to Explain Variance in the VoiceMOS Challenge 2022 Dataset. 4531-4535. doi: 10.21437/interspeech.2022-799
[43] Mingshuai Liu; Shubo Lv; Zihan Zhang; Runduo Han; Xiang Hao; Xianjun Xia (2023). Two-Stage Neural Network for ICASSP 2023 Speech Signal Improvement Challenge. doi: 10.1109/icassp49357.2023.10094827
[44] Shutong Wu, Jiongxiao Wang, Wei Ping, Weili Nie, Chaowei Xiao (2023). Defending against Adversarial Audio via Diffusion Model. doi: 10.48550/arxiv.2303.01507.
[45] Pengxu, Jiang., Cairong, Zou. (2022). Convolution neural network with multiple pooling strategies for speech emotion recognition. 89-92. doi: 10.1109/ISCSIC57216.2022.00029.
[46] Shyamapada, Mukherjee., Neeraj, Shivam., Astha, Gangwal., Lokesh, Khaitan., Amlan, Jyoti, Das. (2019). Spoken Language Recognition Using CNN. 37-41. doi: 10.1109/ICIT48102.2019.00013.
[47] Feng, Ye., Jun, Yang. (2021). A Deep Neural Network Model for Speaker Identification. Applied Sciences, 11(8):3603-. doi: 10.3390/APP11083603.
[48] Taiga, Ishii., Ryo, Ueda., Yusuke, Miyao. (2023). Empirical Analysis of the Inductive Bias of Recurrent Neural Networks by Discrete Fourier Transform of Output Sequences. arXiv.org, abs/2305.09178 doi: 10.48550/arXiv.2305.09178

[49] Jin Wang, Yongsong Zou1, Se-Jung Lim (2023). An Improved Time Feedforward Connections Recurrent Neural Networks. 36(3):2743-2755. doi: 10.32604/iasc.2023.033869.
[50] Haozhe Liu, Wentian Zhang, Bing Li, Haoqian Wu, Nanjun He, Yawen Huang, Yuexiang Li, Bernard Ghanem, Yefeng Zheng (2023). Improving GAN Training via Feature Space Shrinkage. doi: 10.48550/arxiv.2303.01559
[51] Wentian Zhang, Haozhe Liu, Bing Li, Jinheng Xie, Yawen Huang, Yuexiang Li, Yefeng Zheng, Bernard Ghanem (2023). Dynamically Masked Discriminator for Generative Adversarial Networks. doi: 10.48550/arxiv.2306.07716.
[52] Hongjie, Wan., Yuting, Zhou. (2023). ACRCNET: a small audio classification residual convolutional neural network. 12721:1272118-1272118. doi: 10.1117/12.2683351

BIOGRAPHIES OF AUTHORS

Mr. Mandar Pramod Diwakar is a distinguished scholar and educator, holding a B.E. (Computer Engineering) and M.E. (Computer Engineering) from Savitribai Phule Pune University, Pune, Maharashtra, India. Currently pursuing a Ph.D. in Computer Engineering, his research focus lies in Artificial Intelligence within the esteemed confines of Savitribai Phule Pune University, Pune. With an illustrious tenure spanning over 7 years in academia and 2 years in the industry, Mr. Diwakar serves as an Assistant Professor at the Department of Artificial Intelligence & Data Science at Vishwakarma Institute of Information Technology, Pune. His expertise spans a spectrum of domains including Natural Language Processing (NLP), Machine Learning, and Artificial Intelligence (AI). Mr. Diwakar's contributions to academia extend beyond the classroom, with numerous publications gracing both national and international journals and conferences. His research efforts have led to the granting of 8 patents, with 3 more awaiting imminent approval. Additionally, he has authored 3 copyrighted publications, further solidifying his impact on the field of Computer Engineering.

Dr. Brijendra Gupta is a distinguished educator with a profound background in Artificial Intelligence in Bioinformatics. He earned his Doctor of Philosophy (Ph.D.) degree from the esteemed Indian Institute of Technology, Banaras Hindu University (IIT-BHU), following rigorous research and study from 2011 to 2014. Currently, Dr. Gupta holds a pivotal position at Siddhant College of Engineering, Pune, where he contributes his expertise to shaping the academic landscape and fostering the growth of aspiring engineers. Additionally, he serves as an affiliated Ph.D. Guide at Smt. Kashibai Navale College of Engineering and Research Centre, (Savitribai Phule Pune University) guiding and mentoring the next generation of researchers in their pursuit of knowledge and innovation.