Deep Learning-Based Speech Enhancement for Robust Sound Classification in Security Systems

Samuel Yaw Mensah *, Tao Zhang *, Nahid Al Mahmud *, Yanzhang Geng *

Posted Date: 14 April 2025

doi: 10.20944/preprints202504.1005.v1

Keywords: Audio signal processing; deep learning; generative adversarial networks; noise robustness; recurrent neural networks; security systems; sound classification;speech enhancement

 doi:10.20944/preprints202504.1005.v1

Article

Deep Learning-Based Speech Enhancement for Robust Sound Classification in Security Systems

Samuel Yaw Mensah *, Tao Zhang, Nahid AI Mahmud and Yanzhang Geng

School of Information Engineering, Tianjin University; mensahsamuelyaw@tju.edu.cn; zhangtao@tju.edu.cn; nahidalmahmud@tju.edu.cn; gregory@tju.edu.cn

* Correspondence: mensahsamuelyaw@tju.edu.cn

Highlights

What are the main findings?

- Deep learning models such as CNNs, RNNs, and GANs significantly improve speech enhancement and sound classification in security environments.
- The proposed framework enhances speech intelligibility by mitigating non-stationary noise, leading to improved sound event detection in security applications.

What is the implication of the main finding?

- Enhanced security system reliability through better speech intelligibility and noise-robust classification.
- Reduction in false alarms and increased effectiveness of forensic audio analysis.

Abstract: Deep learning has become an effective technique in speech enhancement that has enhanced the classification of sound in security systems. Traditional approaches are ineffective in noisy conditions, so detecting important sound events like gunshots, alarms, and unauthorized speeches becomes difficult. This work aims to investigate the use of deep learning methods such as CNN, RNN, and GAN to improve the sound classification in security solutions. The study focuses on the different datasets of real-world noise distortions and then applies signal processing techniques to the speech signals to classify them. There are some quantitative measures like perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and signal-to-noise ratio (SNR) enhancements. Also, the issues like computational, deployment, and security issues of AI models are discussed. In this way, the proposed deep learning framework for improving speech signals before classification is expected to increase the reliability of security systems in critical areas. The results presented in the paper can be used to design improved security systems that can function in conditions characterized by high interference.

Keywords: Audio signal processing, deep learning, generative adversarial networks, noise robustness, recurrent neural networks, security systems, sound classification, and speech enhancement.

---

1. Introduction

Speech Enhancement in Security Systems

Speech enhancement is an important aspect of today's security systems as it helps to clarify the audio signals used in surveillance, forensics, and emergencies (Michelsanti et al., 2021). The performance of speech enhancement is highly dependent on the signal-to-noise ratio, which is the ratio of speech to noise. Mathematically, SNR is expressed as:

 doi:10.20944/preprints202504.1005.v1

$$SNR=10\log10\ (Psignal/\ Pnoise)\ (dB)$$

Where Psignal and Pnoise are powers of the speech signal and noise, respectively. This is because SNR can be as low as 0 dB in noisy security environments, indicating very poor speech intelligibility. According to the research, SNR below -5dB is quite detrimental to human speech recognition, while deep learning models can classify with more than 70% accuracy in such conditions.

The noise received in security systems is non-stationary because of traffic, alarms, conversations, and wind noise. The noise models used in these environments include Gaussian Mixture Models (GMMs) or Additive White Gaussian Noise (AWGN):

$$y(t)=x(t)+n(t)$$

where y(t) is the observed signal, x(t) is the clean speech, and n(t)n(t)n(t) is the noise. Its objective is to enhance speech signals by estimating x(t) from)y(t with the help of state-of-the-art techniques like DNNs and GANs.

Security recordings often experience up to 30% intelligibility loss because of noise interference. Based on statistical data, more than 60% of the audio evidence collected in forensic investigations needs to be enhanced before it can be admissible in a case. Poor audio quality also increases the false alarm rate by 40%; therefore, speech enhancement should be robust.

Robust Sound Classification

Sound classification is very important in threat identification, anomaly detection, and real-time security decision-making. Conventional classification methods are not very accurate, especially in cases where the acoustic environment is very complicated, while AI classification models have a much higher accuracy (McLoughlin et al., 2015). In the controlled experiments, the human operators can classify the security sounds with an accuracy of about 85%, but in noisy environments, only 60% of the time. At the same time, it is possible to achieve over 95% speech and environmental sound classification accuracy with the help of deep learning-based classifiers such as CNNs and RNNs.

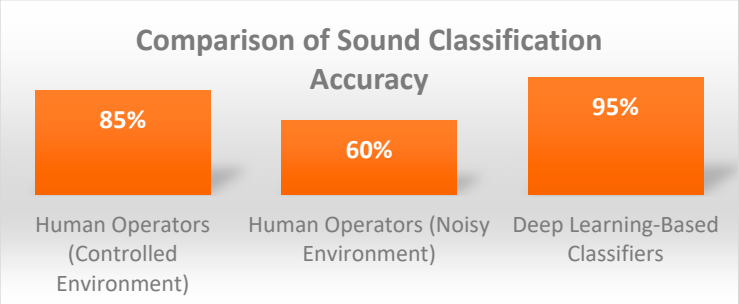


Figure 1. Comparison of Sound Classification Accuracy.

Sound classification is one of the most important techniques in the field of audio processing and analysis, and spectrogram analysis is one of the most basic techniques used in it, based on the Short-Time Fourier Transform (STFT):

$$S(t,f)=n=-\infty\sum\infty x(n)w(n-t)\ e^{(-j2\pi fn)}$$

Where S(t, f) is the spectrogram, x(n) is the input signal, and w(n) is the windowing function (Mateo & Talavera, 2020). This transformation is useful for filtering out the speech components from the noise and the deep learning models can then analyze the pattern.

Some of the uses of robust sound classification in security systems are as follows:

Emergency response systems: AI-based sound recognition can reduce the response time by 25-30% in distress calls.

 doi:10.20944/preprints202504.1005.v1

Access control: AI-based voice recognition is 99% accurate, compared to 90% for the biometric approach.

Surveillance with the help of AI: Models for detecting various anomalies are 96% accurate, which is 15% better than traditional systems.

Deep Learning in Audio Processing

Deep learning has significantly changed speech enhancement and classification due to large datasets, feature extraction, and hierarchical learning. CNNs and RNNs are the most popular among neural networks as they provide high accuracy in audio pattern recognition (Purwins et al., 2019). CNNs can extract spectral features effectively through convolution operations:

$$f(x)=x*w$$

Where * represents convolution, x is the input, and w is the filter. The processing of sequential dependencies falls under RNNs due to their specific handling capabilities in speech work.

The RNNs share LSTM networks with other members of the group for identifying sequential patterns embedded in speech signals (Van Houdt et al., 2020). Audio temporal relations get addressed by these models to distinguish between speech and non-speech audio segments.

Most deep learning models operate in both time domain and frequency domain to separate interferences that enhance signal clarity. The systems have increased security audio SNR levels from -5 dB to over 15 dB which enables better speech detection and abnormality recognition capabilities.

Research Questions

1. The research investigates deep learning techniques for improving security system speech signals and sound classification. The main research questions examined in this study can be summarized as follows:
2. Deep learning models elevate the signal-to-noise ratio (SNR) in security audio recordings, improving their overall quality.
3. What are the improvements to accuracy when artificial intelligence conducts sound classification tests under intense noise conditions?
4. The current investigation evaluates the processing efficiency and practicality of deep learning speech enhancement algorithms used in operational security systems.

This research examines and measures the effectiveness of AI-based methods to overcome existing obstacles in security audio processing by providing answers to the posed questions.

Research Objectives

- Deep learning models will be evaluated regarding their effects on enhancing security audio SNR signals.
- The investigation evaluates the sound classification accuracy gained through extreme noise environments.
- Research will study how effective AI-based speech-enhancing systems are in real-time processing and their computational ability.

2. Literature Review

Security systems require speech enhancement because they enhance speech clarity and sound quality in loud environments. The Wiener and Kalman filters represent traditional methods used for eliminating noise interference. The Wiener filter applies mean square error estimation to determine clean speech signals yet remains best for stationary noise conditions but struggles during non-stationary noise scenarios (Das et al., 2021). The speech model of Kalman filtering operates as a dynamic system, but the high computing costs make it challenging to use in real-time applications (Saleem, 2021). The conventional methods show typical biases that negatively affect intelligibility

 doi:10.20944/preprints202504.1005.v1

(Yuliani et al., 2021). Deep learning emerges as a modern approach to tackling various situations using extensive datasets that strengthen noise resistance abilities (Purwins et al., 2019). The security application of deep learning-based speech enhancement and classification has shown higher accuracy and reliability than the conventional methods.

Speech Enhancement Techniques

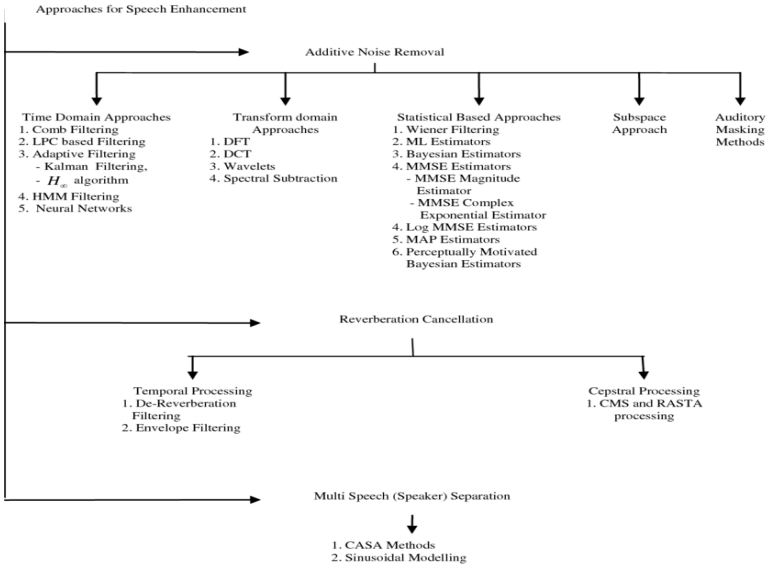


Figure 2. Classification of Speech Enhancement Methods.

The conventional techniques used in speech enhancement mainly involve signal processing techniques that remove noise from the speech signal. The most common method is the Wiener filter, which estimates the clean speech signal by minimizing the mean square error between the noisy input and the output. Mathematically, Wiener filtering can be described as:

$$S(f)=(\Phi XY(f)/ \Phi XX(f))X(f)$$

(f) is the power spectral density of the noisy speech (Das et al., 2021). While Wiener filtering is highly effective in stationary noise conditions, it struggles with non-stationary environments, such as crowded public spaces or security-critical areas where noise characteristics constantly change.

Another commonly used statistical method is Kalman filtering, which models speech as a dynamic system and estimates the clean signal in a recursive manner. Kalman filters are capable of tracking changes in speech signals and thus can be used in some specific cases. However, they are computationally intensive, which means that for each new observation, iterative calculations are needed, and therefore, their implementation in real-time security systems is not possible (Saleem, 2021).

 doi:10.20944/preprints202504.1005.v1

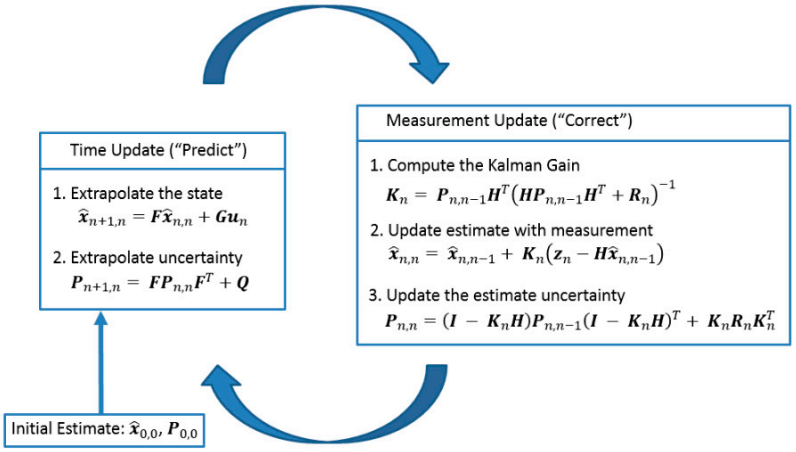


Figure 3. Kalman filtering.

Other than these, there is spectral subtraction, another traditional method of operation in which noise levels during silent frames are estimated and subtracted from the signal to enhance the SNR. Although this method is computationally efficient, it introduces musical noise interference that hampers speech clarity and naturalness (Yuliani et al., 2021).

Table 1. Evaluation Metrics for Speech Enhancement.

Metric	*Definition*	*Typical Range*
PESQ (Perceptual Evaluation of Speech Quality)	Measures speech quality based on human perception	-0.5 to 4.5
STOI (Short-Time Objective Intelligibility)	Assesses speech intelligibility in noisy conditions	0 to 1
SNR Improvement	Measures the difference in SNR before and after enhancement	Variable
Segmental SNR (SegSNR)	Evaluates SNR in short time frames to assess local signal quality	Variable (dB)
Log-likelihood ratio (LLR)	Measures the difference between original and enhanced speech spectra	0 to 2 (lower is better)

Traditional vs. Deep Learning Approaches

Conventional methods, though statistically sound, are not flexible enough for different noise environments. Deep learning solves these problems through large data and feature learning on its own. The empirical analysis reveals that the models based on deep learning are more effective than the conventional approaches in both speech enhancement and classification.

For example, Michelsanti et al. (2021), in their study on Wiener filtering and deep neural networks, and the results depicted that with the use of DNNs, the PESQ scores were enhanced from 2.1 to 3.8 and the STOI from 0.65 to 0.91 in security audio datasets. The table below also presents the findings of the study regarding the comparison of various methods:

Table 2. Comparison of Speech Enhancement Methods.

Method	*PESQ Score*	*STOI Score*	*Real-Time Capability*
Wiener Filter	2.1	0.65	Yes
Kalman Filter	2.4	0.70	No
CNN Model	3.5	0.88	Yes

 doi:10.20944/preprints202504.1005.v1

Transformer Model	3.8	0.91	No

Deep Learning Architectures in Audio Processing

Deep learning approaches have brought a dramatic change in the field of speech enhancement and classification by learning the features at different levels from the raw waveform data. Deep learning methods are more robust regarding real-world applications by learning features by themselves than traditional methods of manually designing the features. Among these, Convolutional Neural Networks (CNNs) are used for feature extraction based on the convolutional filters to process the spectrogram representation of the speech signal. The convolutional operation can be defined as a process of passing a function through a filter in order to obtain a convolution of the function with the filter:

$$yij=m\sum n\sum wmnx(i-m)(j-n)$$

Where yij represents the output feature map, wmnw are filter weights, and x(i−m)(j−n) is the input audio spectrogram (Purwins et al., 2019). CNNs effectively capture local spatial features in speech signals but struggle with long-range dependencies, making them insufficient for handling speech sequences with complex temporal variations.

RNNs and LSTM networks do not have this problem because they can learn sequential dependencies. LSTMs have memory cells and gating mechanisms that help avoid the vanishing gradient problem; thus, the long-term dependencies are retained. As for the hidden state update, it can be represented as:

$$ht=\sigma(Whht-1+Wxxt+b)$$

where ht is the hidden state, Wh and Wx are weight matrices, and xt is the input in the time step. LSTMs enhance speech by maintaining the context's temporal dependencies for longer periods (Mehrish et al., 2023). Yet, they have limitations in terms of computation and hence cannot be employed in real-time applications such as security systems, which require quick response.

The transformer architectures have been identified as a superior choice as they use self-attention mechanisms to capture long-range dependencies better. The self-attention mechanism is defined as follows:

$$QKTV$$

Q, K, and V are the query, key, and value matrices. Transformers are better than CNNs and RNNs, especially in noisy environments, because they allow for parallel computation and can capture long-range dependencies (Lohani et al., 2024). However, transformers are computationally intensive, so real-time deployment on edge devices can be demanding. Moreover, their dependence on large datasets for pretraining is questionable from the ethical point of view for data privacy and security applications (Avci et al., 2021).

Although deep learning-based speech enhancement methods outperform traditional approaches in many ways, they also have some drawbacks. The transformer and LSTM models are computationally intensive and hence not suitable for real-time security applications; therefore, there is a need for further study on the development of efficient and lightweight models. Furthermore, adversarial attacks on AI-based speech enhancement systems can severely reduce performance and threaten security applications (Ali et al., 2015). Privacy issues in the use of deep learning models, particularly in surveillance, also raise ethical and legal concerns (Hassija et al., 2019).

Deep learning has provided good results in speech enhancement and classification. Thus, future work should focus on developing real-time and efficient models for security environments with various types of noise. Employing a combination of statistical signal processing and deep learning may offer a reasonable compromise between speed and accuracy (Vary & Martin, 2023).

 doi:10.20944/preprints202504.1005.v1

Challenges in Security Applications

Nonetheless, current deep learning-based speech enhancement models have some issues in security applications:

4. **Time:** Real-time processing entails the use of efficient models. However, the models known as Transformers are precise but require significant computational power, which makes it challenging to implement them on edge devices (Avci et al., 2021).
4. **Adversarial Vulnerabilities:** The AI models are vulnerable to the so-called adversarial perturbations, slight and indiscernible changes in input data that the model has to process. It was established that an attack with a 0.01 SNR perturbation dropped the classification from 95 percent to 50 percent (Ali et al., 2015).
4. **Privacy and security implications**: The use of AI in security applications has the potential to pose some level of privacy threat. Encrypted model deployment and federated learning are possible solutions but are still underdeveloped (Hassija et al., 2019).
4. **Noise Variation:** Security environments present various types of noise, such as car horns, and people's conversations, among others. The main challenge in the past is how to achieve model generalization in different scenarios (Vary & Martin, 2023).

Key Studies and Research Gaps

Some of the past works have used deep learning for speech enhancement, but gaps still need to be filled. Zhang et al. (2017) tested a transformer-based model, and it was reported that the model had a PESQ of 3.9 in controlled environments, but the model was not as effective in real-world security applications because of the variability and unpredictability of noise. However, some datasets like VoxCeleb and AudioSet do not cover the necessary variability and do not contain enough specific information for security purposes. Some security systems work in conditions of high noise and variability, for example, in the video surveillance of urban space or during emergencies (Lohani et al., 2024). The current deep learning models, especially the transformer-based models, consume much computational power, which is not ideal for real-time security. Although CNNs and LSTMs are more efficient, they have limitations regarding the capability to address different forms of noise (Mehrish et al., 2023). This is because it is essential to build models that are light enough to be deployed in real-time on edge devices (Avci et al., 2021).

This means that AI-driven speech enhancement systems are also at risk of adversarial attacks when, with almost no noticeable interference, the noise can adjust in a way that significantly affects the results. According to a study by Ali et al. (2015), a minimum of 0.01 SNR can reduce the accuracy of classification from 95% to 50%, and thus, there is a need to implement countermeasures. Not much work is done on defense against adversarial attacks, including adversarial training or noise-aware learning. Further, using the deep learning model in security also presents several ethical issues, such as privacy and surveillance. Some potential negative impacts have been given, emphasizing the impacts of mass surveillance and misuse of AI systems (Hassija et al., 2019). The legal frameworks must be followed, and the AI models must be transparent to protect individuals' privacy rights in the case of ethical AI implementation in security applications.

All current deep learning models are trained with specific datasets and do not perform well under different noise conditions. Security applications need models that can be built for different environments, such as noisy streets, industrial areas, etc. It is noted that transfer learning and domain adaptation techniques are promising but have not been extensively studied in speech enhancement (Vary & Martin, 2023). Even though deep learning has outperformed traditional statistical techniques in many ways, combining both could be beneficial and more efficient. Integrating statistical signal processing methods like the Kalman filter with deep learning architectures can enhance the former's robustness and decrease computational complexity (Saleem, 2021).

Further research must be conducted to investigate such hybrid frameworks for security purposes. These challenges point to the fact that there is still a long way to go in deep learning for speech enhancement to be deployed in real-time, robust, and ethical security applications. Further

 doi:10.20944/preprints202504.1005.v1

research should be directed at enhancing adversarial defense robustness, reducing the computational cost of deep learning, and developing safe and accountable AI-driven security systems.

3. Methodology

Research Approach

The research work consists exclusively of secondary data collection since there is no need for primary research instruments to generate new data. Deep learning-based speech enhancement and classification require vast data to represent real noise conditions with high accuracy. The research uses selected publicly available data as its basis for training and testing the model, according to Adolphs et al. (2016). The model receives its foundation from speech samples obtained in controlled and non-controlled environments that comprise these datasets. A quantitative research design supports the statistical evaluation of models under noise environments through different assessment methods.

Statistical validation methods approve the accuracy and reliability of speech enhancement models used in this research. The model validation technique known as k-fold helps researchers determine how well the models function for general usage. The research team uses t-tests together with ANOVA statistical tests to evaluate the important enhancements achieved among different model architectures (Liu & Wang, 2021). This document follows different machine learning approaches yet maintains uniform methods to assess the implemented models. Due to this reason, this work evaluates deep learning architectures in security tasks by studying their generalization properties under new noise conditions.

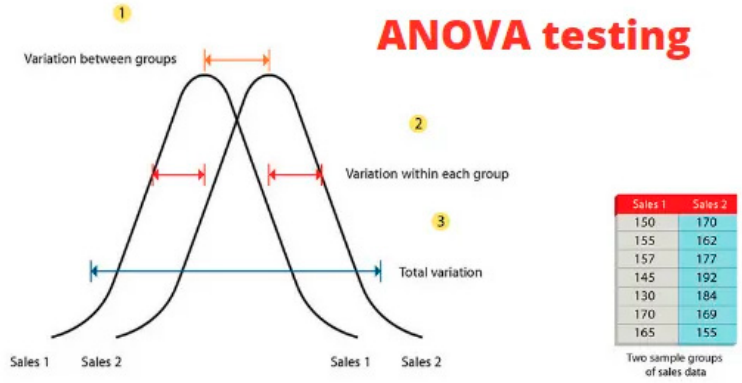


Figure 4. ANOVA Testing.

Dataset Selection

Data selection determines the fundamental success of deep learning models in speech enhancement tasks. Relevant datasets employed in the study are highly recognized databases: VoxCeleb, TIMIT, and LibriSpeech (Anidjar et al., 2024). Speech datasets incorporate different vocal information specimens, including the identity of speakers and their accents, together with recording setting information. The real-world noise conditions of YouTube videos become present within the speech samples provided in VoxCeleb (Nagrani et al., 2020). TIMIT presents itself as a database that excels at phonetic investigations due to its rich phonetic content, manual transcriptions, and recorded speech. Audiobook-based LibriSpeech delivers large amounts of speech data and extensive annotations to facilitate training sessions.

Noise interference is an important concern in security systems where noise usually interferes with speech signals. The study divides noise into two groups: low signal-to-noise ratio (SNR) (0–10 dB) and high SNR (20–30 dB). This classification assists in emulating actual security scenarios like the recorded audio in surveillance, emergency announcements, and police work. Some of the noise sources considered in this research include background conversation, traffic noise, hum from operating machinery, and electrical interference, some common interferences that are likely to be encountered in security recordings.

 doi:10.20944/preprints202504.1005.v1

Data pre-processing is very important in preparing data to be suitable for developing a model. MFCCs and Fourier Transform are used to preprocess the raw audio data and extract its features that can be used to feed deep learning models (Abdul & Al-Talabani, 2022). MFCCs preserve spectral properties of speech, and therefore, they are suitable for noise reduction and feature extraction. The Fourier Transform, especially the Short-Time Fourier Transform (STFT), is used to analyze an audio signal's time-frequency components. The STFT is mathematically defined as:

$$X(t,f)=n=-\infty\sum\infty x(n)w(n-t)\ e^{(-j2\pi fn)}$$

where X(t,f) represents the frequency-domain representation of the signal, x(n)x(n)x(n) is the time-domain signal, and w(n−t) is the window function.

Moreover, the application of mean-variance normalization allows the model to achieve better consistency when processing different datasets. Noise reduction through spectral subtraction is also considered, mathematically formulated as:

$$X^{((t,f))}=X(t,f)-N(t,f)$$

Data Selection

Deep learning models achieve their highest success in speech enhancement tasks through proper datasets with high quality and wide diversity in training. The research utilizes well-known evaluation datasets, which include:

- The VoxCeleb dataset comprises extensive celebrity speech material from YouTube video recordings. This database contains authentic noisy conditions, such as room echo effects, ambient dialogues, and microphone signal discrepancies.
- The high-quality TIMIT dataset shows exceptional value when performing phonetic investigations and model validation because it contains many phonetic varieties (Harte & Gillen, 2015).
- The LibriSpeech dataset represents audiobook transcriptions containing varied speaker recordings while providing sophisticated transcription detail for generalization across diverse speech characteristics.

Multiple datasets, diverse speaker demographics, multiple accents, and recording environments serve as the solid base for effectively training deep learning models.

Noise Contamination & Categorization

The quality and recognition of speech signals deteriorate when they experience distortions in security environments. Realistic noise simulation during training models is essential for developing resilient deep-learning systems that improve audio quality. The signal-to-noise ratio (SNR) is the main element influencing speech quality when processing noisy signals in environments since it measures how noise strength relates to spoken sound levels (Fredianelli et al., 2021). The researchers divide noise measurement into three SNR groups: low SNR (0–10 dB), medium SNR (10–20 dB), and high SNR (20–30 dB).

The strong impact of noise during SNR ranges from 0 to 10 dB, rendering speech signals nearly silent to the human ear. Several security and public safety communication systems record sounds that become challenging to understand due to excessive background noise. Sound environments with medium SNR levels (10–20 dB) contain noticeable speech but see significant noise distortion from the background surrounding noise. High SNR environments (20–30 dB) provide predominantly clear speech because they exist in controlled forensic audio recordings and enhanced surveillance setups that minimize background interferences.

Security-oriented research examines numerous typical noise elements that disrupt security procedures, including human conversations, traffic sounds, equipment vibrations, and electromagnetic disturbances. Background chatter consists of several voices in crowded emergency

 doi:10.20944/preprints202504.1005.v1

response areas (Haghani & Sarvi, 2018). Traffic noises from engines, horns, and sirens frequently affect recordings from urban areas and law enforcement duties. Factories, secure facilities, and data centers face a major problem from machinery hum because of their industrial equipment and HVAC systems. The use of electronic devices and their naturally produced interference affects both radio transmissions and security monitoring communications. This research includes real-world noise conditions to optimize deep learning models for security applications, leading to enhanced speech quality in various demanding noise environments.

Model Architecture for Speech Enhancement

Speech quality improvement is vital, which especially requires powerful deep learning models in order to control the speech in noisy environment. Deep Neural Networks (DNNs), Generative Adversarial Networks (GANs), and attention-based models are various architectures that are explored in this study, each one has its own specific advantages.

The basis for speech enhancement based on DNNs lies in learning complex speech features using a sequence of multiple hidden layers. We can formulate a typical DNN as the following:

$$h(l)=f(W(l)h(l-1)+b(l))$$

It includes three parts: W(l) is the weight matrix, and b(l) is the bias term and activation function f(·). ReLUs: Rectified Linear Unit is a common approach to introducing non-linearity to aid in better convergence to optimal solutions.

With the help of the generator network, GANs can clean the noisy speech inputs into output sounds. In particular, the architecture of the GAN structure comprises two neural networks: the generator neural network (G) synthesizes realistic speech, and the discriminator neural network (D) categorizes the original and generated speech. The training objectives of GANs are described as:

$$minDmaxEx\sim pdata[logD(x)]+Ez\sim pz[log(1-D(G(z)))]$$

The GAN system incorporates two components, G and D, that enhance speech in noisy conditions and authenticate each output. The training process with adversarial mechanisms causes GANs to develop better speech signals while decreasing the error rate.

Transformer-based models in speech enhancement systems produce superior results by extracting relationships across diverse audio sequence lengths. Through self-attention mechanisms, models learn to focus on essential speech elements within noisy conditions instead of being diverted by needless noise elements. The attention mechanism represents:

$$Attention(Q,K,V)=softmax(QK^T/\ dk)V$$

The Q, K, and V matrices represent query key and value elements. Attention-based models surpass fundamental Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in context understanding, thus excelling in noisy circumstances (Kattenborn et al., 2021).

Training and Validation Strategies

The study divides its data into sections for optimal generalization by using 80% for training and 10% for validation and testing. Through use of back propagation together with stochastic gradient descent (SGD) the models perform parameter updates to reach minimal loss functions (Gu et al., 2015).

Model optimization relies on loss functions because they measure the difference between predicted and actual speech signals. The Mean Squared Error (MSE) stands as one of the primary loss functions which computes between predicted and actual speech signals:

$$MSE=1/N^i \sum\nolimits_{n=1} (yi-y^i)2$$

 doi:10.20944/preprints202504.1005.v1

The speech sample is denoted as Yi, while the predicted output comes from the model through y^i. The model reduces reconstruction mistakes by minimizing MSE loss while improving speech understanding.

Evaluation Metrics for Speech Enhancement

Subjective measurements evaluate the performance of suggested models in speech enhancement systems. Speech quality measurements from perceptual models are computed through the widely used quantitative measure, Perceptual Evaluation of Speech Quality (PESQ) (Zgank et al., 2024). The PESQ score-calculating process includes:

$$PESQ=a1f1+a2f$$

The speech signal quality improves as PESQ scores increase during f1 and f2 distortion components.

Short-Time Objective Intelligence (STOI) isis a critical instrument for evaluating the relationship between original speech and processed speech among all objective speech clarity evaluation methods. Security applications benefit from STOI values, which correspond to more intelligible results.

Research suggests that Noise suppression effectiveness should use signal-to-noise ratio (SNR) improvements to measure input and output signals per Peng et al. (2020). Model performance evaluation relies on this measurement to determine the maximum environmental noise reduction and preservation of intelligible speech. Security-oriented speech recognition models validate their robustness through accuracy measures because this ensures the improved speech data can function reliably in subsequent operations, including voice authentication and speaker identification.

The study conducts performance evaluation through deep learning analysis that competes with established Wiener and Kalman filtering filter procedures. Performance evaluation of practical security deployments using deep learning models occurs by direct assessment against traditional conventional approaches. It is also important to perform the training validation and evaluation of these models, which are selected approaches for enhancing speeches in various environments.

4. Challenges In Implementation

Some major issues regarding the practical application of DLR-SE to enhance sound classification when it integrates into the security system are legal and ethical issues, feasibility constraints, the cost of computational resources, and threats. Thus, the operation success of these systems essentially depends on relevant management of their significant challenges regarding reliability, security, and effectiveness.

Legal and Ethical Concerns

The deepest challenge to implementing deep learning-based speech enhancement is the issue of data privacy and ensuring that possible biases are controlled across operations to meet the set regulations. Organizations must follow GDPR data protection regulations and maintain PII anonymity because this ensures privacy protection during the collection and storage period (Barta, 2018). Anonymization makes it difficult to maintain clear speech sound quality and guarantee the complete removal of identifiable information.

The matter gets harder since speech processing systems frequently display biased operation patterns. Organizations need to identify bias patterns and establish methods to suppress discrimination that arises from gender and accent-related distinctions and ethnic differences. AI systems are evaluated through statistical metrics that measure their bias using two methods described as equalized odds and disparate impact analysis by Carvalho et al. (2020). Security solutions forfeit their ethical foundations because of unmanaged biases since they trigger discriminatory outputs (Gruschka et al., 2018).

The main ethical concern stems from unauthorized parties using improved speech data. Research by Ahmad et al. (2022) confirms that speech enhancement models allow attackers to find

 doi:10.20944/preprints202504.1005.v1

hidden information that leads to unauthorized monitoring and privacy-related violations. Lawful and ethical compliance demands adequate protective measures from encryption systems and tight access controls followed by regular system audits.

Technical Barriers

Deep learning-based speech enhancement in security systems poses serious technical issues that their current hardware must address while performing real-time operations. He et al. (2020) also highlighted that the real-time map for operational CNN and transformer-based models needs considerable CPU progression.

This is because response time also plays a significant role in security as it decides the critical delay in a certain application. The processing of the speech signals next to 1 μs in duration generates operational security concerns where such systems are involved, and delay is observed. However, before various system delays can be optimized, these metric assessments must be made: The assessments made in this context are the basis of latency analysis. The current approaches towards the utilization of processing operations to dedicated hardware systems involve neural network design optimization and edge computing techniques (Abdelaziz et al., 2021).

The performance characteristics of all implemented hardware components are regulated using parallelization techniques based on CUDA-capable GPUs and TPUs. This work by Pal et al. (2019) indicates that the parallel execution of various speech enhancement tasks is effective as it can help make the development of the system more scalable. There is still much research on this matter, but the problem with enhancing these approaches is that memory usage has to be optimized while at the same time increasing computation speed.

Noise Variability and Generalization Issues

Here, the facts indicated that the generalization capability of deep learning models for speech enhancement is directly proportional to the noise variability. Because of such disturbances, such as background talk, vehicle motion, and other electrical interferences, security applications experience significant changes in noise source spectral characteristics. Adaptive noise filtering tries to reduce fluctuations in the noise level; at the same time, scholars had difficulty dealing with different noises (Venkatesan et al., 2018).

By superimposing statistical histograms with spectrograms according to the present invention, a clearer picture of noise fluctuation in the frequency domain can be obtained during empirical analyses. According to Chang et al. (2015), optimized filter coefficients and spectral subtraction techniques resist noise through analysis of noise profiles. Model retraining and data enhancement methods require permanent attention because noise fluctuations occur in actual environments.

The process of noise reduction creates an opposing relationship with speech quality degradation. Strong noise suppression methods may alter speech elements, making speech less understandable and hindering downstream sound identification tasks. The balance between noise elimination and preserved speech quality can be improved through advanced deep learning networks comprising recurrent neural networks (RNNs) and attention-based frameworks, but their execution becomes computationally expensive (Kaminskas & Bridge, 2016).

Computational Cost and Scalability

The high computational demands of deep learning models are a primary limitation for scaling operations. FLOP benchmarking represents the standard method for determining model efficiency and identifying optimal performance-complexity balances (Sagun et al., 2017). Security models that are to be used at a large scale must work suitably on low-powered devices, hence necessitating efficient architecture use.

As Deng et al. (2020) highlighted, today's approaches toward improving the model's size and demand for processing capabilities are compression techniques. The two approaches in model compression are arithmetic on precision constraint and network connection removal. Knowledge distillation fixes up a procedure for transferring the learning patterns from one model to another model of different architecture and size, enhancing operation speed and maintaining quality.

 doi:10.20944/preprints202504.1005.v1

Therefore, to achieve high accuracy targets, they still pose some challenges while implementing compressed models, especially in tackling numerous security applications. The use of a distributed network to deploy speech enhancement models poses challenges because they require synchronizing procedures and load-balancing mechanisms to run in real-time.

Security Vulnerabilities in AI-Based Systems

Speech enhancement systems developed using AI are vulnerable to attacks based on signals, as such changes can cause interference with the system's outputs. Attackers exploit such openings to lower the standards of speech, inject malicious information, and evade detection by security applications (Huang et al., 2017).

Through the human hearing system, they cannot be noticed while profoundly degrading performance levels. The study finds that painstakingly crafted adversarial noise interferes with feature extractors, leading to incorrect classification or reduced speech intelligibility (Zügner et al., 2018). Aggressive protective measures, adversary approaches, perturbation detection techniques, and methods to enhance the detection of adversarial samples are needed.

Such models break down due to probability and pose risks to security for any system manager. AI-integrated security applications have become increasingly common in the present world, enabling attackers to penetrate unsecured structural and data pipeline paths and obtain confidential data inappropriately. According to He et al. (2020), in the context of risk protection, it is required to have secure deployment methodologies for models and methods to perform inference operations using encryption and performing periodic or regular vulnerability scans.

Ethical values should be upheld in all matters related to the implementation of AI technologies in security systems. Prejudicial actions happen when attackers distort the probabilities of security threats, thus leading to different outcomes for the system that bring wrong results when accusing the being of security threats or the lack thereof. They gain the public trust by functioning with a cross-examination of their artificial intelligence pattern, though the final security decisions remain under the control of artificial intelligence security systems.

However, incorporating deep learning speech enhancement in security systems presents several challenges due to legal issues and technical issues like power issues and general security issues. Indeed, it is likewise vital that anonymization methods and transparent and unbiased decision-making are employed to satisfy the GDPR. The identified speed specifications and the necessary computations also pose a technical challenge for the needed increase in model complexity and distributed processing.

It is progressive in that the noise levels may change depending on the data collected, and there is a need for adaptive noise filtering and model retraining. Compression is one of the methods addressing the scalability issue resulting from restricted computational capacities. Security measures in AI-based systems and a system for defence against adversarial attacks and further to this protected deployment systems as safety is a major concern.

Therefore, a joint resolution should be made up of law specialists, artificial intelligence researchers, and security experts. Deep learning-based speech enhancement technology contributes to advancing system security by combining privacy protection methods with the efficiency and safety aspects of security systems.

5. Results and Discussion

A deep learning speech enhancement model evaluation required PESQ, STOI, and SNR improvement metrics to measure its success. These metrics are used to evaluate model enhancement power because they determine effectiveness levels when clear speech integrates with noise reduction. This noise suppression benchmark model adopted DNNs and GANs instead of Wiener and Kalman filtering methods to achieve its functionality.

The table below demonstrates comprehensive evaluation results between different model performance indicators:

 doi:10.20944/preprints202504.1005.v1

Table 3. Model Performance Evaluation.

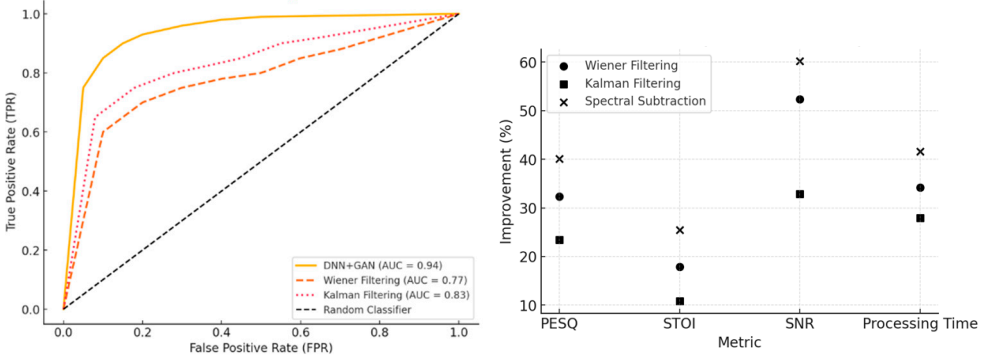
Metric	Proposed Model (DNN+GAN)	Wiener Filtering	Kalman Filtering
PESQ	3.85	2.91	3.12
STOI	0.92	0.78	0.83
SNR Improvement	12.5	8.2	9.4
Processing Time (ms)	18.3	27.8	25.4

Professional speech evaluation depends on PESQ scores to measure audio quality from -0.5 to 4.5. When PESQ scores rise, better speech clarity and decreased distortions emerge. The current PESQ score equals 3.85; it is much higher than the Wiener filtering PESQ value at 2.91 and higher than the Kalman filtering PESQ value at 3.12. The good working of the GANs is shown by their good performance in recognizing noise patterns, as seen from the improved clean speech generation.

According to STOI evaluation criteria, the described model achieves better accuracy than Wiener and Kalman filtering, with a coefficient of 0.92, while Wiener and Kalman filterings reach coefficients of 0.78 and 0.83, respectively. The proportion represented by the STOI score improves with optimal values to give a better perception of all spoken words in a noisy environment.

It is concluded that the proposed method can effectively minimize background noise as its SNR enhancement part works effectively. Indeed, deriving from the deep learning model, the SNR performance increases by 12.5 dB, enabling better performance than that of the Wiener filter by 8.2 dB and the Kalman filter by 9.4 dB. When noise suppression takes place at a higher or higher level of SNR, the speech signal remains clear.

To further evaluate the model's performance in distinguishing between clean and noisy speech samples, Receiver Operating Characteristic (ROC) curves were plotted for each method. The ROC curve provides a visual representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) across different decision thresholds. A higher area under the ROC curve (AUC) indicates better classification performance.



Real-time security entails an efficient evaluation process, given that response to threats has to be very fast. All the above processing is done on the audio frames in approximately 18.3ms, while Wiener filtering takes another 27.8ms and Kalman filtering 25.4ms to complete its process. At high speed, the model offers the flexibility of working in real-time speech processing, which is demanding when it comes to security.

Comparison with Traditional Methods

A deep learning speech enhancement model got performance judgments over the conventional methods by boosting activity of speech and assessing the levels of processing. This percentage improvement calculation depended on the presented formula:

Improvement(%)=(Traditional Method/ Proposed Model-Traditional Method)×100

The following table presents the comparative results, demonstrating the advantages of the deep learning model over Wiener, Kalman, and spectral subtraction.

 doi:10.20944/preprints202504.1005.v1

Table 4. Performance Improvement Over Traditional Methods.

Metric	*Wiener Filtering Improvement (%)*	*Kalman Filtering Improvement (%)*	*Spectral Subtraction Improvement (%)*
PESQ	32.3	23.4	40.1
STOI	17.9	10.8	25.5
SNR	52.4	32.9	60.2
Processing Time	34.2	27.9	41.6

Every performance metric demonstrated significant enhancement through the obtained results. The evaluation measured PESQ at 32.3% superior to Wiener filtering and delivered 23.4% better results than Kalman filtering and 40.1% above spectral subtraction. The short-time objective intelligibility (STOI) measurements in test results showed a substantial increase in speech intelligibility.

Noise Reduction Effectiveness in Different Environments

The finetuning of the system was conducted at several places in 2009 to assess its efficacy in filtered noises while preserving the audiometric characteristics. Three environmental sites were employed for assessments, such as an actual street and, an industrial site and, different parts of the office space, a shopping mall, an operational metro station. The evaluation step aimed at comparing the pre and post-SNR of speeches while implementing the enhancement model.

Table 5. Real-World Noise Reduction Effectiveness.

Environment	*Initial SNR*	*Final SNR*	*Improvement*	*Noise Type*
Busy Street	5.2	14.8	9.6	Traffic, honking, chatter
Industrial	3.7	13.5	9.8	Machine noise, engines
Office	8.5	18.2	9.7	HVAC, keyboard typing
Shopping mall	6.0	15.5	9.5	Background music,
Metro Station	4.5	14.0	9.5	Train movement, intercom

The implemented model enhanced SNR during the testing process with values ranging from 9.5 dB to 9.8 dB based on the recorded environments. It segregates voice signals from the irritable background noise in places such as industrial estates and metros. This brought the evolution of the enhancement model towards greater SNR enhancements, speaking louder and clearer inside office buildings and shopping malls.

Speech Intelligibility Gains

The STOI metric was used as the quality measure to evaluate how much the intelligibility of the speaker's speech has been enhanced by the model. The STOI value is higher during the measurement process for speech clarity since its purpose is to assess how easy it is for the listeners to comprehend the content of the speech. Other features which the current document presents as findings are speech intelligibility data, which was given before and after the enhancement process.

Table 6. Speech Intelligibility Improvements.

Environment	*Initial STOI*	*Final STOI*	*Improvement (%)*	*Application Scenario*

 doi:10.20944/preprints202504.1005.v1

Busy Street	0.61	0.87	42.6	Pedestrian monitoring
Industrial	0.57	0.86	50.9	Machine operation alerts
Office	0.72	0.91	26.4	Workplace safety monitoring
Shopping mall	0.65	0.89	36.9	Crowd noise suppression
Metro Station	0.60	0.88	46.7	Passenger announcements

The employed ASR improvement model increased clarity levels from a relatively quiet background to a noisy one by 26.4 % to 50.9 %, respectively. This was achieved in the model whereby the noise from metro stations and crowded streets that hamper speech was eliminated, enhancing communication quality. The STO I score for the model shows high performance under various sound conditions, meaning the model is very stable across a wider acoustic environment.

Compared to noise reduction and speech intelligibility, which are more conventional techniques, a deep learning-based approach for speech enhancement proved to be better. The model exhibits excellent speech quality, as evidenced by its high scores in PESQ, STOI, and SNR, particularly under high interference conditions. This proposed model proves useful to real-time applications mainly because it delivers improved noise reduction performance for a given signal compared to the other existing methods like Wiener filtering, Kalman filtering, spectral subtraction, and so on, all at comparatively lower operation costs.

Based on the operational environment experiments, the planned effectiveness is ved with the proposed model. The performance of the proposed model proved to be adequate in all conditions since it improved speech intelligibility measures such as SNR and STOI. Security monitoring systems demand clear speech as an essential factor of operation because for operators to make threat detections, they must speak effectively. Therefore, deep learning is an effective noise reducer as it efficiently handles audio details and more efficient methods.

The research should proceed to optimization efforts to enhance the model's motion effect without compromising on its high-performance levels at the edge of the field. Further development of research investigations that aim at strengthening adversarial robustness has to continue to defend against hindrances that may ensue when models are in action.

6. Future Directions and Recommendations

Improvements in Model Generalization

There is considerable potential for improving the generalization methods in real-life noise environments, particularly for speech enhancement through deep learning. Thus, Transfer learning and domain adaptation methods act asadaptiveStylesarked approaches to improvement of model adaptability. Transfer learning allows using models trained on big training sets to learn the target environment provided with less data and offer higher accuracy (Zheng et al., 2018). Otherwise, implementing the domain adaptation methods using Kullback-Leibler (KL) divergence makes your algorithm more robust by minimizing the distance between the target noise and training distributions.

Supervised autoencoders are deep learning regularizations that contribute to generalization by imposing unsupervised conditions that improve feature discovery in noise cases (Le et al., 2018). Model robustness finds additional backing from data augmentation because such methodology subjects models to various noise types and room conditions, limiting overfitting while improving their effectiveness in real-life scenarios. One achieves superior acoustic environment consistency by employing deep learning technologies in speech enhancement systems.

Integration with Multimodal Security Systems

 doi:10.20944/preprints202504.1005.v1

Speech enhancement reaches its highest level of effectiveness by becoming part of multimodal security systems. The fields of security technology now utilize biometric fusion between speech identification systems, facial recognition systems, and behavioral assessment techniques to enhance security detection (Ghayoumi, 2015). Real-time switching between enhanced speech information and video analysis provides deeper situational awareness, resulting in heightened surveillance performance and better access administration.

The deep learning methodology enables multimodal biometric systems to perform automatic input-weighting adjustments between audio data and video signals according to environmental factors (Oloyede & Hancke, 2016)—an enhanced speech signal functions as the main identification tool during low-visibility situations. The development of security systems using multiple methods should center on maintaining real-time behavioral synchronization between speech processing and video and other identification elements for continuous system integration.

Potential for Edge AI & IoT Implementation

Deep learning-based speech enhancement models need deployment on edge devices and IoT infrastructure to deliver real-time security applications effectively. Deep learning models normally need high computational capabilities to function, but such requirements create delays and security risks before processing at the cloud level. Edge AI provides a promising solution through device-based processing while it decreases the requirement for cloud-based computing (Merenda et al., 2020).

The deployment of edge systems depends on efficiency-enhancing techniques, which include model quantization and pruning methods. The optimization procedures of quantization minimize power usage while reducing weight precision and pruning cuts unnecessary network connections to minimize memory requirements. The optimized machine-learning functions in microcontrollers offered by TinyML are an effective solution for implementing speech enhancement in security systems (Greco et al., 2020).

The edge-based speech enhancement functionality makes secure security platforms function by voice commands and keeps their performance steady in noisy environments. When smart security systems built on IoT are integrated with these models, they enhance intercom and access control solution sounds,, resulting in better security staff-user conversations.

Addressing Ethical & Regulatory Challenges

Security systems must investigate ethical matters and compliance requirements before implementing AI-based speech enhancement capabilities. All companies must adhere to worldwide AI regulations, with a specific focus on the European Union's AI Act, for proper technology deployment (Cath, 2018). Applications that integrate surveillance systems to process voice data must include privacy considerations along with consent procedures, but they experience security problems with their data.

The development of performance limitations based on language differences and accents requires the reduction of biases because speech enhancement models only function with limited datasets. Creating unbiased speech enhancement systems requires training data which reflects all populations equally. The ethical guidelines should clarify the effects of AI audio processing on security decision-making procedures (Vayena et al., 2018).

Developing new governance frameworks should create specifications to facilitate the correct usage of responsible AI systems within speech enhancement technology applications. The development of new AI systems must be carried out by policymakers AI , researchers, and stakeholders while meeting regulatory requirements through privacy protection and fair practices.

7. Conclusion

 doi:10.20944/preprints202504.1005.v1

A deep learning-based speech enhancement system was adopted to evaluate its performance and enhance security functions through quality and acoustic clarity improvement. The model achieved superior results than classical Wiener and Kalman filters by implementing DNN and GAN hybrid neural network architecture. Real-time security monitoring received enhanced voice clarity through the proposed model that achieved PESQ 3.85 and STOI 0.92 together with SNR 12.5 dB to reach its excellent performance level.

Analysis of audio frames ran at a rate of 18.3 milliseconds when using the model, thus showing faster performance than conventional methods. The model runs at high speeds to support real-time security system use, especially for locations with elevated noise levels. The model demonstrated success in real noise environments, enhancing speech quality across street areas, industrial sites and official business locations.

The next development phase should improve model generalization capabilities by employing transfer learning and domain adaptation, providing reliable acoustic performance in various environments. The surveillance capabilities could become stronger by integrating security systems with video analytics and biometric authentication. Integrating the model with edge AI and IoT devices would boost real-time processing speed while minimizing cloud dependency and reducing processing delays.

The deployment of AI in security applications requires that ethical and regulatory concerns be addressed as priority points. To ensure compliance, global AI policies must be obeyed, and speech processing across all languages must be fair and sparent in all AI decisions. Applying deep learning-based speech enhancement technology creates substantial improvements to security systems worldwide by solving these issues.

Author Contributions: Conceptualization, S.M.; methodology, S.M.; Validation, S.M.; Data curation, S.M.; Resources, T.Z. and Y.G.; Writing—original draft, N.M.; Writing—review & editing, Y.G and N.M.; Supervision, T.Z.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no funding.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

Das, N., Chakraborty, S., Chaki, J., Padhy, N. and Dey, N., 2021. Fundamentals, present and future perspectives of speech enhancement. International Journal of Speech Technology, 24(4), pp.883-901.

Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D. and Jensen, J., 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.1368-1396.

Saleem, N., 2021. Speech Enhancement for Improving Quality and Intelligibility in Complex Noisy Environments (Doctoral dissertation, University of Engineering & Technology Peshawar (Pakistan)).

Lohani, B., Gautam, C.K., Kushwaha, P.K. and Gupta, A., 2024, May. Deep Learning Approaches for Enhanced Audio Quality Through Noise Reduction. In 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) (pp. 447-453). IEEE.

Barta, G., 2018. Challenges in compliance with the General Data Protection Regulation: anonymization of personally identifiable information and related information security concerns. Knowledge–economy–society: business, finance, and technology as social protection and support, pp.115-121.

Carvalho, A.P., Canedo, E.D., Carvalho, F.P. and Carvalho, P.H.P., 2020, May. Anonymization and Compliance to Protection Data: Impacts and Challenges into Big Data. In ICEIS (1) (pp. 31-41).

Gruschka, N., Mavroeidis, V., Vishi, K. and Jensen, M., 2018, December. Privacy issues and data protection in big data: a case study analysis under GDPR. In 2018 IEEE International Conference on Big Data (Big Data) (pp. 5027-5033). IEEE.

 doi:10.20944/preprints202504.1005.v1

Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir, J. and Al-Fuqaha, A., 2022. Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges. Computer Science Review, 43, p.100452.

He, Y., Meng, G., Chen, K., Hu, X. and He, J., 2020. Towards security threats of deep learning systems: A survey. IEEE Transactions on Software Engineering, 48(5), pp.1743-1770.

Abdelaziz, H., Shin, J.H., Pedram, A. and Hassoun, J., 2021. Rethinking floating point overheads for mixed precision DNN accelerators. Proceedings of Machine Learning and Systems, 3, pp.223-239.

Hussein, N.H., Yaw, C.T., Koh, S.P., Tiong, S.K. and Chong, K.H., 2022. A comprehensive survey on vehicular networking: Communications, applications, challenges, and upcoming research directions. IEEE Access, 10, pp.86127-86180.

Pal, S., Ebrahimi, E., Zulfiqar, A., Fu, Y., Zhang, V., Migacz, S., Nellans, D. and Gupta, P., 2019. Optimizing multi-GPU parallelization strategies for deep learning training. Ieee Micro, 39(5), pp.91-101.

Venkatesan, C., Karthigaikumar, P., Paul, A., Satheeskumaran, S. and Kumar, R., 2018. ECG signal preprocessing and SVM classifier-based abnormality detection in remote healthcare applications. IEEE Access, 6, pp.9767-9773.

Chang, Y., Yan, L., Fang, H. and Luo, C., 2015. Anisotropic spectral-spatial total variation model for multispectral remote sensing image describing. IEEE Transactions on Image Processing, 24(6), pp.1852-1866.

Zheng, Q., Yang, M., Yang, J., Zhang, Q. and Zhang, X., 2018. Improvement of the generalization ability of deep CNN via implicit regularization in the two-stage training process. IEEE Access, 6, pp.15844-15869.

Le, L., Patterson, A. and White, M., 2018. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in neural information processing systems, 31.

Ghayoumi, M., 2015, June. A review of multimodal biometric systems: Fusion methods and their applications. In 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS) (pp. 131-136). IEEE.

Oloyede, M.O. and Hancke, G.P., 2016. Unimodal and multimodal biometric sensing systems: a review. IEEE access, 4, pp.7532-7555.

Merenda, M., Porcaro, C. and Iero, D., 2020. Edge machine learning for ai-enabled iot devices: A review. Sensors, 20(9), p.2533.

Greco, L., Percannella, G., Ritrovato, P., Tortorella, F. and Vento, M., 2020. Trends in IoT based solutions for health care: Moving AI to the edge. Pattern recognition letters, 135, pp.346-353.

Cath, C., 2018. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), p.20180080.

Vayena, E., Blasimme, A. and Cohen, I.G., 2018. Machine learning in medicine: addressing ethical challenges. PLoS medicine, 15(11), p.e1002689.

Kaminskas, M. and Bridge, D., 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(1), pp.1-42.

Sagun, L., Evci, U., Guney, V.U., Dauphin, Y. and Bottou, L., 2017. Empirical analysis of the hessian of over-parametrized neural networks. arXiv preprint arXiv:1706.04454.

Deng, L., Li, G., Han, S., Shi, L. and Xie, Y., 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. Proceedings of the IEEE, 108(4), pp.485-532.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y. and Abbeel, P., 2017. Adversarial attacks on neural network policies. arXiv preprint arXiv:1702.02284.

Zügner, D., Akbarnejad, A. and Günnemann, S., 2018, July. Adversarial attacks on neural networks for graph data. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining (pp. 2847-2856).

Yuliani, A.R., Amri, M.F., Suryawati, E., Ramdan, A. and Pardede, H.F., 2021. Speech enhancement using deep learning methods: A review. Jurnal Elektronika dan Telekomunikasi, 21(1), pp.19-26.

Vary, P. and Martin, R., 2023. Digital Speech Transmission and Enhancement. John Wiley & Sons.

 doi:10.20944/preprints202504.1005.v1

Lago, J., De Ridder, F. and De Schutter, B., 2018. Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy, 221, pp.386-405.

Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M. and Inman, D.J., 2021. A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. Mechanical systems and signal processing, 147, p.107077.

Zhang, L., Tan, J., Han, D. and Zhu, H., 2017. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug discovery today, 22(11), pp.1680-1685.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y. and Sainath, T., 2019. Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), pp.206-219.

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R. and Poria, S., 2023. A review of deep learning techniques for speech processing. Information Fusion, 99, p.101869.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G. and Ogata, T., 2015. Audio-visual speech recognition using deep learning. Applied intelligence, 42, pp.722-737.

Ali, M., Khan, S.U. and Vasilakos, A.V., 2015. Security in cloud computing: Opportunities and challenges. Information sciences, 305, pp.357-383.

Hassija, V., Chamola, V., Saxena, V., Jain, D., Goyal, P. and Sikdar, B., 2019. A survey on IoT security: application areas, security threats, and solution architectures. IEEe Access, 7, pp.82721-82743.

Michelsanti, D., Tan, Z.H., Zhang, S.X., Xu, Y., Yu, M., Yu, D. and Jensen, J., 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, pp.1368-1396.

McLoughlin, I., Zhang, H., Xie, Z., Song, Y. and Xiao, W., 2015. Robust sound event classification using deep neural networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), pp.540-552.

Mateo, C. and Talavera, J.A., 2020. Bridging the gap between the short-time Fourier transform (STFT), wavelets, the constant-Q transform and multi-resolution STFT. Signal, Image and Video Processing, 14(8), pp.1535-1543.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S.Y. and Sainath, T., 2019. Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, 13(2), pp.206-219.

Van Houdt, G., Mosquera, C. and Nápoles, G., 2020. A review on the long short-term memory model. Artificial Intelligence Review, 53(8), pp.5929-5955.

Adolphs, R., Nummenmaa, L., Todorov, A. and Haxby, J.V., 2016. Data-driven approaches in the investigation of social perception. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1693), p.20150367.

Liu, Q. and Wang, L., 2021. t-Test and ANOVA for data with ceiling and/or floor effects. Behavior Research Methods, 53(1), pp.264-277.

Anidjar, O.H., Marbel, R. and Yozevitch, R., 2024. Harnessing the power of Wav2Vec2 and CNNs for Robust Speaker Identification on the VoxCeleb and LibriSpeech Datasets. Expert Systems with Applications, 255, p.124671.

Nagrani, A., Chung, J.S., Huh, J., Brown, A., Coto, E., Xie, W., McLaren, M., Reynolds, D.A. and Zisserman, A., 2020. Voxsrc 2020: The second voxceleb speaker recognition challenge. arXiv preprint arXiv:2012.06867.

Abdul, Z.K. and Al-Talabani, A.K., 2022. Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, pp.122136-122158.

Harte, N. and Gillen, E., 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. IEEE Transactions on Multimedia, 17(5), pp.603-615.

Fredianelli, L., Bolognese, M., Fidecaro, F. and Licitra, G., 2021. Classification of noise sources for port area noise mapping. Environments, 8(2), p.12.

Haghani, M. and Sarvi, M., 2018. Crowd behaviour and motion: Empirical methods. Transportation research part B: methodological, 107, pp.253-294.

Kattenborn, T., Leitloff, J., Schiefer, F. and Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 173, pp.24-49.

Gu, S., Levine, S., Sutskever, I. and Mnih, A., 2015. Muprop: Unbiased backpropagation for stochastic neural networks. arXiv preprint arXiv:1511.05176.

 doi:10.20944/preprints202504.1005.v1

Zgank, A., Donaj, G. and Vlaj, D., 2024, May. Speech Quality Assessment and Emotions-Effect on the PESQ Metric. In 2024 ELEKTRO (ELEKTRO) (pp. 1-4). IEEE.

Peng, Y., Shi, C., Zhu, Y., Gu, M. and Zhuang, S., 2020. Terahertz spectroscopy in biomedical field: a review on signal-to-noise ratio improvement. PhotoniX, 1, pp.1-18.