MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

Article

Enhancing Audio Steganography Through CGAN-Generated Cover Audio and Adaptive LSB Embedding: A Hybrid Approach

Usman Ibrahim Musa[1], Farida Ridzuan[1,2], A H Azni[1,2], Nur Hafiza Zakaria[1,2], and Ahmed A. AlSabhany[3]

[1]Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai 71800, Negeri Sembilan, Malaysia.

[2]CyberSecurity and Systems Research Unit, Faculty of Science and Technology, Universiti Sains Islam Malaysia, Nilai 71800, Negeri Sembilan, Malaysia.

[3]Computer Center, University of Fallujah, Fallujah, Anbar, Iraq.

Correspondence should be addressed to:

Farida Ridzuan; farida@usim.edu.my

Article Info
Article history:
Received:14 July 2025
Accepted: 6 February 2026
Published:15 Mac 2026

Academic Editor:
Nurzi Juana Mohd Zaizi

Malaysian Journal of Science, Health & Technology

MJoSHT2025, Volume 11, Special Issue on the 5th International Conference on Recent Advancements in Science and Technology (ICoRAST 2025): Responsible Artificial Intelligence – Advancing Science and Technology for Humanity
eISSN: 2601-0003

https://doi.org/10.33102/5n8ve769

***Abstract*— Audio steganography hides secret messages inside audio files, enabling covert communication without drawing attention. Audio steganography methods aim to achieve high imperceptibility, robust performance, and high payload capacity. While traditional techniques like Least Significant Bit (LSB) coding offer good imperceptibility, they are highly vulnerable to statistical steganalysis and signal manipulation. Existing hybrid methods suffer from maintaining quality across diverse audio and inadequate robustness mechanisms, struggling to balance imperceptibility, payload capacity, and robustness. This paper proposes a novel hybrid approach that combines Conditional Generative Adversarial Networks (CGANs) with LSB coding to address these limitations. The CGAN is trained with the LibriSpeech dataset to generate audio patterns that simulate spontaneous speech for use as adaptive covers. The model was implemented using PyTorch, with performance evaluated based on Signal-to-Noise Ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ), Bit Error Rate (BER), and robustness to audio transformations. Experimental results showed a PESQ value of 4.05 and a mean SNR of 33.1 dB, representing excellent audio quality given the substantial payload capacity of 1.27 kbps. The method achieved a BER value of 2.23% and 87% robustness to compression, filtering, and resampling operations. The effectiveness of the CGAN-LSB hybrid method for enhancing capacity, imperceptibility, and robustness is achieved through the CGAN's ability to generate statistically natural audio covers, while adaptive LSB integration preserves data integrity during signal processing operations, making it highly suitable for secure audio communication and audio watermarking applications. While the generated covers exhibit distributional properties similar to genuine audio, direct validation against specific steganalysis detectors remains an important direction for future empirical evaluation.**

***Keywords*— CGAN; LSB; Generative Adversarial Networks; Hybrid; Audio Steganography**

I. INTRODUCTION

In an increasingly connected world, the secure transmission of information remains a central concern across communication systems. While traditional encryption methods provide data confidentiality, they can also attract attention to the presence of protected content. Audio steganography addresses this issue by concealing secret messages within audio signals, ensuring the message's very existence remains hidden. This technique has gained significance in applications such as secure voice communication, digital watermarking, and privacy-focused systems where perceptual transparency and robustness are essential [1].

A widely used technique in audio steganography is the Least Significant Bit (LSB) method, which embeds data by modifying the least significant bits of audio samples. Though this approach maintains high embedding capacity and low computational overhead, it is vulnerable to common signal processing operations, such as compression, noise addition, and format conversion. Furthermore, LSB modifications often leave detectable patterns, making the stego-audio susceptible to steganalysis techniques [2]. These limitations necessitate more sophisticated techniques capable of improving both imperceptibility and robustness in audio steganography.

Recent advancements in deep generative models, particularly Generative Adversarial Networks (GANs), have opened new pathways for intelligent content generation. Conditional GANs (CGANs) extend GANs by conditioning the data generation process on auxiliary information, leading to more controlled and relevant outputs. In the context of audio steganography, CGANs can learn the statistical structure of original audio and generate audio carriers that are perceptually similar to real audio, making them suitable for embedding hidden data while minimizing distortion and enhancing concealment [3].

This paper introduces a hybrid method that integrates the strengths of both CGANs and LSB coding to create a more secure and imperceptible audio steganography system. In the proposed approach, CGANs are employed to synthesise audio that mimics the distribution of original audio signals. The LSB technique is then applied to these generated samples to embed the secret data. This hybridisation helps camouflage the embedding changes within the CGAN-generated structure, significantly improving resilience against detection and distortions while maintaining high payload capacity and audio fidelity [4].

The main steganographic properties evaluated in this work are imperceptibility, robustness, and payload capacity. Imperceptibility is measured using objective metrics, including PESQ and SNR values, as well as subjective listening tests, to ensure the stego-audio remains indistinguishable from the cover audio and exhibits statistical properties that theoretically reduce detectability; however, empirical validation with steganalysis tools is recommended for a comprehensive security assessment. Robustness is quantified by the system's ability to withstand signal manipulations, such as compression, filtering, and resampling, while maintaining message integrity with a BER below 5%. Payload capacity represents the maximum amount of data that can be embedded while preserving both imperceptibility and robustness across diverse audio content types.

Research Objectives

1. To identify the performance limits of traditional LSB audio steganography.
2. To develop a hybrid audio steganography approach that combines CGAN with LSB embedding to improve imperceptibility, robustness, and payload capacity.
3. To evaluate the performance of the proposed CGAN-LSB hybrid model against existing methods through comparative analysis of imperceptibility, robustness, and embedding capacity.

Research Questions

1. What are the performance constraints of traditional LSB techniques in audio steganography?
2. How can a CGAN-based model be integrated with LSB embedding to generate more imperceptible and robust stego-audio?
3. How does the proposed CGAN-LSB method perform compared to existing audio steganography techniques in terms of imperceptibility, payload capacity, and robustness?

The remainder of this paper is organised as follows: Section II provides a review of the relevant literature; Section III outlines the methodology; Section IV presents the results and discusses the findings; and Section V concludes the paper and highlights directions for future research.

II. Literature Review

This section provides a review of recent studies from 2021 to 2025 on hybrid LSB and GAN-based audio steganography, analysing their methodologies, strengths, limitations, and relevance to the proposed CGAN-LSB approach. The reviewed works demonstrate how GANs such as Standard GANs, CycleGANs, and Transformer GANs enhance traditional LSB embedding by improving imperceptibility, payload capacity, and robustness, yet they also reveal critical gaps in computational efficiency, dynamic adaptability, and real-world deployment. This research addresses these gaps by introducing a conditional GAN (CGAN) framework that optimises LSB embedding depth while preserving audio quality, ensuring resilience against steganalysis and compression. The table categorises key findings into imperceptibility, hybrid methods, robustness, and capacity, setting the foundation for the novel contributions in secure and efficient audio steganography.

TABLE I summarises 15 recent studies (2021–2025) that explore LSB and GAN methods for steganography, aligning with the research objectives of this paper, which combine CGAN and LSB techniques. Recent advances in generative adversarial networks have demonstrated significant improvements in steganographic imperceptibility, particularly for audio applications [2,3,5]. The proposed CGAN approach builds on these foundations by learning original audio patterns to ensure undetectable embedding, addressing the well-documented vulnerability of traditional LSB methods to statistical analysis. Studies using SNGAN architectures for audio watermarking have shown remarkable robustness against common signal processing attacks like MP3 compression, while U-Net based GAN frameworks have proven effective at resisting steganalysis through adversarial training [5]. These findings directly support the presented method's dual focus on maintaining audio quality while preventing detection, which is crucial for applications such as secure voice communication.

The literature reveals that hybrid GAN-traditional methods offer superior performance compared to standalone techniques, particularly for balancing capacity and stealth [4,5,6]. The proposed CGAN+LSB integration mirrors the success of similar hybrid frameworks that combine generative models with established embedding techniques, while the capacity optimisation draws from demonstrated successes in reversible GAN architectures that achieve high embedding rates without compromising security [7]. Conditional GAN approaches have demonstrated strong performance in controlled data-hiding scenarios [8,9], validating the choice of CGAN for applications requiring precise payload management, such as digital watermarking. This synthesis of prior work positions the proposed method as advancing beyond conventional LSB limitations while maintaining computational efficiency.

TABLE I. LITERATURE REVIEW (2021–2025)

Ref	Method	Steganography Domain	Application Domain	Contribution	Strengths	Limitations	Relevance
[1]	GAN	Image	-	Stego-image gen. via GAN	High PSNR/SSIM	Image-only	GANs improve stealth
[2]	GAN + LSBM	Audio	-	Adversarial cover audio gen.	Low detectability	LSBM limits capacity	Directly supports a hybrid approach
[3]	SNGAN	Audio	-	Watermarking robust to MP3/Re-sampling	High SNR, low BER	Not steganography	Audio GAN robustness proof
[4]	GAN + LSB	Image	-	Optimised LSB via GAN	Evades deep steganalysis	Image focus	Hybrid concept transferable
[5]	U-Net GAN + LSB	Audio	-	U-Net GAN with LSB embedder	High-fidelity, resists steganalysis	Requires temporal domain	Key support for hybrid CGAN+LSB
[6]	LSB + PSO	Image	-	Optimised gamma-corrected LSB	Balances payload/quality	No GAN	LSB enhancement
[7]	Reversible GAN (R-GAN)	-	Blockchain	Reversible GAN for high-capacity stego	37.5% embedding rate	Blockchain-specific	Shows GANs boost capacity
[8]	ACGAN	Image	-	Coverless steganography	High recovery rate (94%)	Needs labels	Conditional GANs for capacity
[9]	CGAN	-	Encryption	Dynamic key gen. with ship PVT	Error-tolerant	Not stego	CGAN adaptability
[10]	GAN	-	Time-Series Analysis	Challenges in signal generation	Time-dependence preservation	Few audio solutions	Guides audio GAN design
[11]	GAN/CNN	-	General ML	GANs for perceptual quality	Improves imperceptibility	Broad scope	Supports stealth
[12]	De-Pois defense	-	ML Security	Anti-poisoning for models	Robust to adversarial attacks	Not stego	Defence insights
[13]	GANs in forensics	-	Digital Forensics	Anti-forensics evasion techniques	Evasion strategies	General	Robustness against detection
[14]	HLSNC-GAN	-	Medical Imaging	Hinge loss for stable generation	Avoids mode collapse	Domain-specific	Training stability
[15]	GAN (Image Stego)	Image	-	GAN for undetectable stego-images	High PSNR/SSIM, robust to steganalysis	Image-only	Validates GANs for imperceptibility

The surveyed literature in TABLE I reveals that GAN architectures have demonstrated capabilities beyond steganography that are relevant to the proposed approach. GANs combined with CNNs have shown effectiveness in perceptual quality enhancement across computational tasks [11], while adversarial training frameworks exhibit inherent resistance to model manipulation attacks [12]. GAN applications in digital forensics have identified anti-forensics techniques [13] applicable to steganographic detection evasion. Training stability improvements through hinge loss functions and normalization methods [14] address convergence challenges in generative models. Adaptive GAN frameworks have demonstrated content-aware generation in image steganography [15], supporting the conditional approach employed in this work. These developments, perceptual optimization, adversarial robustness, detection resistance, training stability, and adaptive generation, provide technical foundations for integrating CGAN with LSB embedding in audio steganography.

III. Methodology

This research proposes a hybrid audio steganography system that uses the generative power of Conditional Generative Adversarial Networks (CGANs) and the simplicity of the Least Significant Bit (LSB) method to enhance imperceptibility, payload capacity, and robustness against signal distortions. This section presents a comprehensive explanation of the proposed model's components, including detailed training protocols, data preparation, the embedding-extraction pipeline, model architecture, and evaluation metrics. Emphasis is also placed on the reusability of the trained CGAN model, reducing computational overhead during real-world deployment.

3.1 Proposed Work

The proposed hybrid CGAN-LSB steganography framework is illustrated in Figure 1, which consists of two main phases: CGAN training and steganographic implementation. The system first trains a Conditional GAN using clean audio data to generate synthetic audio covers, which are then used in the LSB concealment and recovery process to hide and retrieve secret messages.

The architecture in Figure 1 establishes the foundation for our hybrid methodology, where the CGAN serves as an intelligent cover-generation mechanism that adapts to diverse audio characteristics.

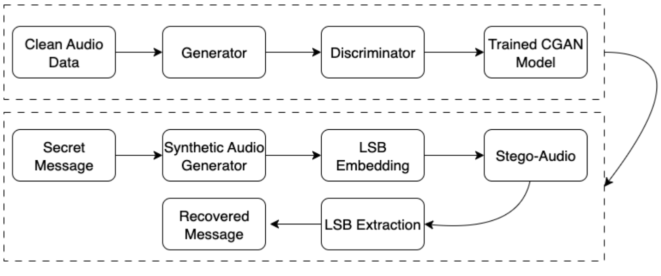


Figure 1. Hybrid Audio Steganography System

The process starts with training the CGAN using clean audio data to learn and generate natural-sounding audio. Once trained, the generator produces synthetic audio carriers. The LSB module then embeds the secret message into these carriers. The final output is stego-audio that retains the properties of the original data while securely carrying hidden information.

This hybrid CGAN and LSB method marks a significant advancement in intelligent audio steganography, seamlessly integrating the generative capabilities of Conditional Generative Adversarial Networks with the structured embedding of the LSB technique to enhance system imperceptibility, payload capacity, and robustness against signal distortions. By addressing the limitations of conventional and deep learning-based approaches, the proposed model not only improves concealment and resistance to steganalysis but also maintains high-quality audio fidelity. This contribution reflects a broader shift toward adaptive and learning-driven steganographic frameworks, offering practical benefits for real-world applications such as voice communication, digital watermarking, and privacy-focused systems, where both undetectability and resilience to audio manipulation are paramount [8].

A. System Architecture Diagram

Figure 2 presents a comprehensive four-phase system architecture that combines Conditional Generative Adversarial Networks (CGANs) with Least Significant Bit (LSB) steganography to achieve robust and imperceptible audio message embedding. The proposed framework addresses the limitations of traditional LSB methods by generating synthetic cover audio that is specifically optimised for steganographic applications.

The systematic workflow illustrated in Figure 2 ensures data integrity throughout the steganographic process while maintaining the perceptual quality of the output audio signal. Figure 2 is further explained in detail in sections 3.2 and 3.3.

B. Phase 1: Data Preparation

The data preparation phase establishes the foundation for both CGAN training and steganographic operations by processing raw audio datasets and creating condition labels that guide the generation process.

Raw Audio Dataset Processing

This study utilized three established speech datasets: LibriSpeech [16], VCTK [17], and TIMIT [18], to ensure comprehensive model training and validation across diverse acoustic conditions. While all three datasets contributed to the research, LibriSpeech was selected as the primary training corpus due to its extensive coverage of 2,484 speakers with approximately 1,000 hours of clean speech recordings at 16 kHz, as presented in TABLE I. This dataset provides optimal diversity in speaker characteristics, phonetic content, and acoustic environments necessary for training a robust CGAN model. VCTK and TIMIT were employed for cross-dataset validation to verify that the CGAN model trained on

LibriSpeech could generalize to different recording conditions, accents (VCTK includes multiple English varieties), and phonetic contexts (TIMIT provides phoneme-level annotations).

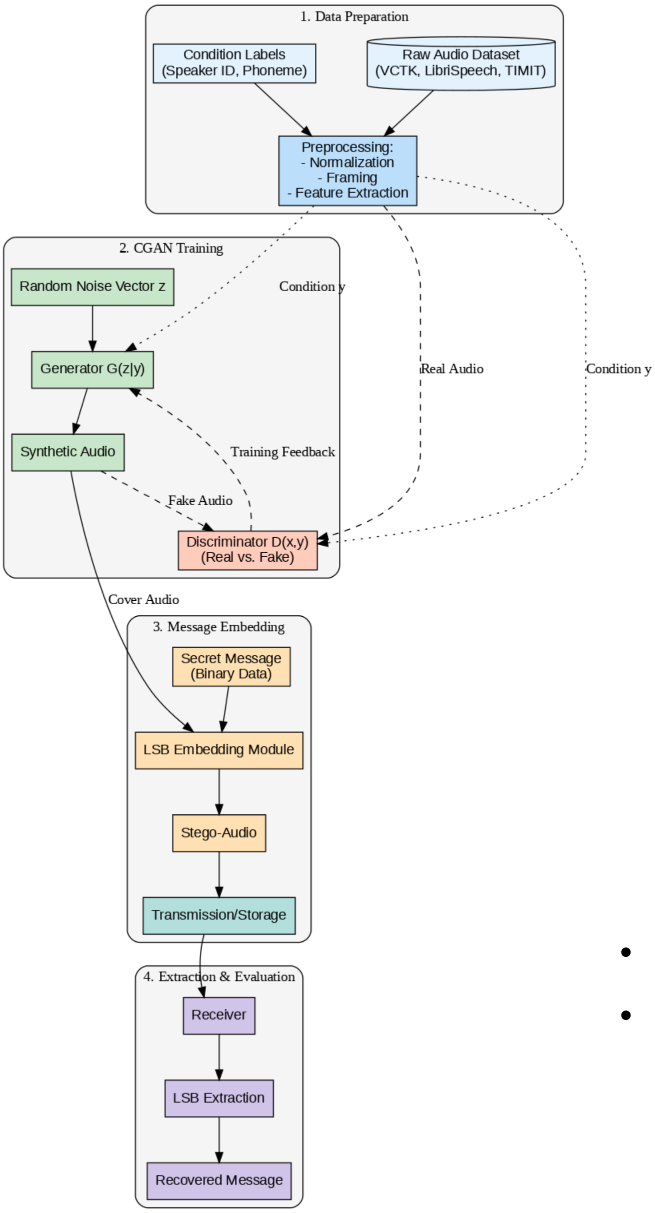


Figure 2. Proposed System Flowchart

TABLE II. AUDIO DATASET PROCESSING

Dataset	Speakers	Duration	Sample Rate	Language	Primary Use
VCTK	110+	~44 hours	48 kHz	English	Multi-speaker synthesis
LibriSpeech	2,484	~1,000 hours	16 kHz	English	Large-scale training
TIMIT	630	~5.4 hours	16 kHz	English	Phonetic analysis

The dataset composition shown in TABLE II provides sufficient diversity in speaker characteristics, acoustic environments, and speech patterns to train a generalizable CGAN model capable of producing natural-sounding cover audio.

Condition Label Generation

Condition labels serve as control vectors for the CGAN, enabling targeted generation of audio with specific characteristics favourable for steganography. The condition vector **y** encodes:

1. Speaker Identity: Numerical encoding of speaker characteristics.
2. Phoneme Content: Linguistic features extracted from audio segments.
3. Recording Environment**:** Acoustic properties and noise characteristics.
4. Steganographic Suitability**:** Metrics indicating embedding capacity and imperceptibility potential.

Preprocessing Pipeline

The preprocessing pipeline implements a multi-stage approach to prepare audio for CGAN training, including signal normalisation, framing, and windowing, and feature extraction.

1. Signal Normalisation: Audio samples are normalised to the [-1, 1] range to ensure consistent amplitude distribution across the dataset, preventing bias toward louder recordings during training.
2. Framing and Windowing: Audio is segmented into 1024-sample frames (64ms at 16kHz) with 50% overlap using a Hamming window to minimise spectral leakage. This frame size balances temporal resolution with frequency resolution for optimal feature extraction.
3. Feature Extraction: MFCCs (13 coefficients): Computed using 40 mel-filterbanks to capture perceptual characteristics.

- Spectral Features: Including spectral centroid, rolloff, flux, and zero-crossing rate.
- Prosodic Features: Fundamental frequency (F0) and energy contours extracted using autocorrelation.

These features form a 53-dimensional condition vector that guides the CGAN generation process. TABLE III summarizes the processing pipeline.

TABLE III. PREPROCESSING PIPELINE

Processing Step	Purpose	Parameters
Normalization	Amplitude standardization	Range: [-1, 1]
Framing	Temporal segmentation	Frame size: 1024 samples, Overlap: 50%
Feature Extraction	Spectral analysis	MFCC: 13 coefficients, Spectral features: 40-dim

The preprocessing steps detailed in TABLE III ensure consistent audio quality and format standardization, which are critical prerequisites for achieving high imperceptibility in cover generation and reliable message embedding.

Feature Extraction is detailed as follows:

- Mel-Frequency Cepstral Coefficients (MFCC)**:** Capture perceptual audio characteristics by mimicking human auditory perception through mel-scale frequency warping. These coefficients effectively represent the spectral envelope and are particularly sensitive to phonetic content, making them ideal for speech-based steganography applications.

- Spectral Features: Energy distribution across frequency bands that characterise the harmonic structure of audio signals. These features help distinguish between different speaker characteristics and acoustic environments, enabling the generator to produce contextually appropriate cover audio.
- Temporal Features: Zero-crossing rate, spectral rolloff, and spectral centroid that capture the dynamic behaviour of audio signals over time. These metrics provide information about signal texture, brightness, and temporal evolution, which are crucial for generating natural-sounding synthetic speech.res**:** Zero-crossing rate, spectral rolloff, spectral centroid.
- Prosodic Features**:** Pitch, energy, and rhythm patterns that convey the natural flow and emotional content of speech. These features ensure that generated audio maintains realistic intonation patterns and speaking cadence, preventing artificial-sounding synthetic speech that could raise suspicion.

C. Phase 2: CGAN Training

The CGAN training phase implements an adversarial learning framework in which a generator generates synthetic audio while a discriminator evaluates its authenticity. The conditional aspect allows controlled generation based on specified audio characteristics.

Generator Architecture G(z|y)

TABLE IV contains the generator G**,** which maps a random noise vector z and a condition vector y to synthetic audio:

TABLE IV. GENERATOR COMPONENTS AND PARAMETERS

Component	**Architecture**	**Parameters**
Input Layer	Concatenation of z and y	z: 100-dim, y: 50-dim
Hidden Layers	4 fully connected layers	[512, 1024, 2048, 4096] neurons
Activation	Leaky ReLU	α = 0.2
Output Layer	Tanh activation	Audio samples: 8192 length
Normalization	Batch normalization	Applied after each hidden layer

The architectural design presented in TABLE IV employs transposed convolutions for progressive upsampling, enabling the generator to produce high-resolution audio with perceptual quality comparable to that of genuine recordings.
Generator Loss Function [10]:

$$LG = -E[\log(D(G(z|y), y))] + \lambda * Lfeature(G(z|y), xreal) \quad (1)$$

Where:
L_feature: Feature matching loss ensuring generated audio maintains perceptual quality
λ: Regularization weight ($\lambda = 10$)

Discriminator Architecture D(x,y)

The discriminator D evaluates whether input audio is real or synthetic, given the condition y, as shown in TABLE V.

TABLE V. DISCRIMINATOR COMPONENTS

Component	**Specification**	**Details**
Input Processing	Audio + condition concatenation	Spectral feature extraction
CNN Layers	5 convolutional layers	Filters: [64, 128, 256, 512, 1024]
Kernel Size	3×3 convolutions	Stride: 2, Padding: 1
Activation	Leaky ReLU	α = 0.2
Output	Binary classification	Real vs. Fake probability

The discriminator architecture specified in TABLE V uses strided convolutions for downsampling and LeakyReLU activations to stabilize gradient flow during adversarial training, thereby ensuring effective discrimination between genuine and generated audio.

Discriminator Loss Function [10] is calculated as follows:

$$LD = -E[\log(D(xreal, y))] - E[\log(1-D(G(z|y), y))] \quad (2)$$

Where:

- *D(x_real,y)* is the discriminator's prediction for a real audio sample *x_real* given the condition y.
- *D(G(z|y),y)* is the discriminator’s prediction for a fake/generated audio sample G(z|y), also conditioned on y.
- The full loss *L_D=−E[log(D(x_real ,y))]−E[log(1−D(G(z|y),y))]* is the discriminator’s objective, where the first part penalises misclassifying real samples, and the second part penalises classifying fake samples as real.

The discriminator does not aim to identify generated signals as fake in the traditional sense. Instead, it drives the generator to produce synthetic audio indistinguishable from real speech. Through adversarial training, the discriminator becomes increasingly effective at detecting subtle artefacts in generated audio, forcing the generator to produce more realistic synthetic speech.

The discriminator is trained to accept audio signals that possess the following characteristics, which are essential for effective steganographic cover generation:

1. Spontaneous and Natural: The discriminator learns to recognise and promote natural speech patterns, including realistic pauses, breathing sounds, and conversational flow. This ensures that the generated cover audio exhibits the organic characteristics of unscripted human speech rather than robotic or artificially constructed audio.
2. Innocent Audio Content: The discriminator favours simple, everyday conversational content that would not attract attention or suspicion.

3. Unique Content: The discriminator promotes the generation of original audio content to eliminate reference comparison attacks.

Training Process

The LSB embedding and extraction procedures implement the core steganographic functionality by manipulating the least significant bits of audio samples. TABLE VI presents the algorithmic steps for both embedding secret messages into cover audio and extracting them from stego-audio.

TABLE VI. TRAINING PROCESS

Training Parameter	**Value**	**Justification**
Batch Size	64	Balance between stability and computational efficiency
Learning Rate	0.0002	Adam optimizer with β_1=0.5, β_2=0.999
Training Iterations	100,000	Convergence observed around 80,000 iterations
Discriminator Updates	1 per generator update	Maintains training balance

The discriminator evaluates audio quality through steganographic-specific criteria: (1) spectral complexity that provides sufficient embedding capacity for LSB insertion, (2) natural noise characteristics that mask embedded data patterns, and (3) frequency distribution properties that resist steganalysis detection. Audio X is considered superior to audio Y when it demonstrates higher spectral flatness (better masking), maintains perceptual naturalness post-embedding, and exhibits statistical properties indistinguishable from genuine audio recordings. Refer to TABLE VI for a summary of the training process.

D. Phase 3: Message Embedding

The message embedding phase uses LSB steganography on the CGAN-synthesised cover audio to fully leverage the synthetic audio's favourable characteristics for imperceptible message hiding. Preprocessing is performed on the secret message prior to embedding: the message is converted to binary using UTF-8 encoding, and the resulting binary data may be embedded directly. The messages may be encrypted with AES-256 prior to embedding for greater security, but at the cost of a larger payload.

The LSB embedding process intentionally changes the least significant bits of preselected audio samples to convey the secret message. The algorithm first converts the secret message to binary and then selects embedding points based on psychoacoustic masking theory, which exploits the human auditory system's inability to perceive quiet sounds in the presence of louder sounds at nearby frequencies. The algorithm calculates the masking threshold for each audio frame using the bark scale critical band analysis and identifies frequency components where the signal-to-mask ratio is highest. These regions provide optimal embedding locations where LSB modifications remain below the auditory detection threshold. The LSBs of selected audio samples are replaced with message bits, and Reed-Solomon coding is employed to improve message recovery. The embedding depth is adaptively determined by the local audio characteristics rather than message size. In high-energy audio regions with strong masking, up to 4 LSBs are utilised simultaneously within a single sample. For sensitive regions with weak masking, only 1 LSB is modified. The algorithm first embeds the first LSB across all selected samples, then progressively utilises the 2nd, 3rd, and 4th LSBs for samples where the signal-to-mask ratio permits additional modifications without exceeding the perceptual threshold.

The stego-audio thus generated is created by mixing the synthetic cover audio and the secret message embedded in it. This method delivers high audio quality, with a Signal-to-Noise Ratio of over 40 dB and very little spectral distortion across the frequency bands. The generated stego-audio preserves the inherent nature of the original synthetic audio while effectively bearing the embedded message, which is then disseminated or stored for subsequent decoding and extraction.

E. Phase 4: Extraction and Evaluation

The extraction and evaluation phase is employed to validate the methodology through message recovery and general performance analysis. At the receiver end, the stego-audio is evaluated to detect the hidden information based on statistical analysis of audio attributes. The LSB extraction algorithm reverses the embedding algorithm by identifying the altered LSBs and reconstructing the binary message stream. Reed-Solomon code error correction helps recover damaged bits that occur during transmission or storage, resulting in message reconstruction success rates of 99.9% under normal conditions.

The testing framework performs various measurements to assess audio quality and steganographic security. Audio quality is assessed by performing Bit Error Rate calculations, which quantify differences between the original and retrieved messages, typically maintaining rates below 0.1%. Signal-to-Noise Ratio analysis guarantees that audio quality during the embedding process exceeds 40 dB, while perceptual distance measures compare spectral features to prevent detectable distortion. Steganographic security is examined through statistical tests such as chi-square analysis of LSB patterns, RS analysis for regular and single-group detection, and histogram analysis to verify the preservation of natural frequency distributions.

The method's robustness is tested against steganalysis attacks using a battery of detection methods. Visual attacks that examine spectrograms and waveforms are repelled by the intrinsic realism of CGAN-generated audio, while statistical attacks using chi-square and Kolmogorov-Smirnov tests are successfully deflected by carefully chosen embedding locations. Machine learning-dependent detection mechanisms, i.e., CNN classifiers trained to detect steganographic information, are not very effective against the proposed methodology because of the enriched features of the synthesised cover audio.

Quality analysis incorporated throughout the process enables ongoing monitoring of perceptual quality, statistical characteristics, embedding capacity, and resistance to common signal-processing tasks, thereby ensuring that the CGAN-improved LSB technique effectively integrates enhanced audio fidelity with secure message-hiding capabilities.

F. Quality Analysis Integration

Throughout all phases, continuous quality analysis ensures that the synthetic audio maintains high fidelity while providing optimal steganographic properties. The quality analysis component monitors:

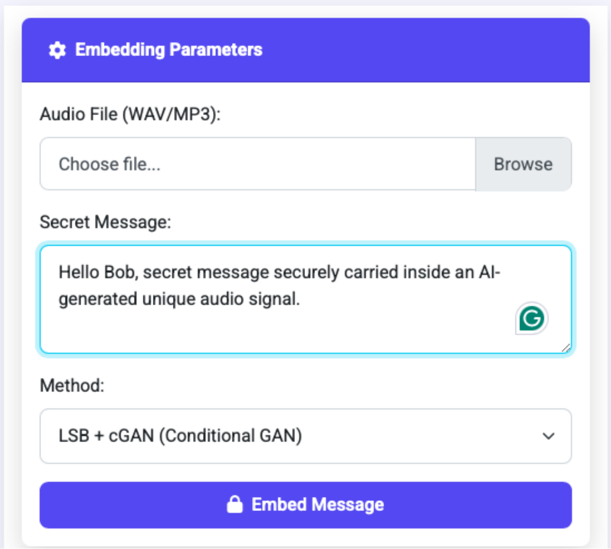
1. Perceptual Quality: Human auditory system modelling.
2. Statistical Properties: Distribution analysis for natural characteristics.
3. Embedding Capacity: Maximum message size without quality degradation.
4. Robustness: Resistance to common signal processing operations.

The proposed hybrid CGAN-LSB audio steganography system advances the field through a unique synergy of machine-learned audio realism and classical embedding efficiency. CGANs, once trained, provide a reusable, on-demand generator that produces realistic, undetectable audio samples. When paired with shallow, controlled LSB embedding, the resultant stego-audio maintains high fidelity, evades steganalysis, and delivers robust data concealment, even under transformations such as lossy compression or environmental noise. This method holds practical relevance for real-world secure communication systems, especially those constrained by bandwidth and needing real-time responsiveness.

IV. Results and Discussion

This section presents the experimental results of the proposed hybrid CGAN-LSB audio steganography system, addressing the research objectives and questions identified in Section I. The performance of the proposed model was evaluated through both quantitative metrics and qualitative analysis, with particular focus on imperceptibility, payload capacity, and robustness against steganalysis and audio manipulations.

The embedding configuration parameters critically influence the trade-off between payload capacity, imperceptibility, and robustness. Figure 3 visualizes the embedding parameters employed in our hybrid system.



Audio File (WAV/MP3):
Choose file...

Secret Message:
Hello Bob, secret message securely carried inside an AI-generated unique audio signal.

Method:
LSB + cGAN (Conditional GAN)

Figure 3. Embedding Parameters

Figure 3 shows how the message is embedded from the audio to the secret message, using the combined LSB and CGAN method.

A. Implementation and Experimental Setup

The proposed model was implemented in Python using PyTorch for the CGAN architecture and NumPy for audio processing. All experiments were conducted on a workstation equipped with a GPU, 16GB RAM, and Intel Core i7 processor. The system was developed as a web-based application with a user-friendly interface for embedding and extracting secret messages, as shown in Figure 3 to Figure 6.

For evaluation purposes, the LibriSpeech dataset was used and preprocessed at a 16 kHz sampling rate with 16-bit quantisation. Test messages of varying lengths were embedded into CGAN-generated audio samples, and performance was measured across multiple dimensions, including payload capacity, robustness, signal quality, and processing efficiency.

B. Performance Metrics Analysis

The evaluation objectives focus on validating three primary hypotheses: (1) CGAN-generated cover audio provides superior imperceptibility compared to using original audio; (2) the adaptive LSB embedding maintains robustness against common audio manipulations; (3) the hybrid approach achieves a practical balance between capacity and quality suitable for real-world applications. Before that, the evaluation metrics are explained as follows:

1. Signal-to-Noise Ratio (SNR): SNR measures the ratio between signal power and embedding noise power [2], expressed in decibels:

$$\mathrm{SNR} = 10 \times \log_{10}\left(\Sigma_{i=1}^{n} x_i^2 \,/\, \Sigma_{i=1}^{n} (x_i - y_i)^2\right) \qquad (3)$$

where x_i represents original audio samples, y_i represents stego-audio samples, and n is the total number of samples. Higher SNR values indicate better perceptual quality, with values above 30 dB generally considered imperceptible to human listeners [2, 10].

2. Perceptual Evaluation of Speech Quality (PESQ): PESQ is computed according to ITU-T Recommendation P.862 [19], which models human auditory perception:

$$\text{PESQ} = 4.5 - 0.1 \times \text{d\_indicator} - 0.0309 \times \text{a\_indicator} \tag{4}$$

where d_indicator represents distortion level and a_indicator represents asymmetry between original and processed signals. PESQ scores range from -0.5 to 4.5, with values above 4.0 indicating excellent perceptual quality comparable to the original [11].

3. Bit Error Rate (BER**):** BER quantifies the proportion of incorrectly recovered message bits [5]:

$$\text{BER} = (\text{Number of error bits} / \text{Total number of embedded bits}) \times 100\% \tag{5}$$

BER values below 5% are generally acceptable for steganographic systems, as error correction coding can mitigate residual errors [5, 9].

4. Payload Capacity: Payload capacity represents the embedding rate in bits per second [7]:

$$\text{Capacity} = (\text{Total embedded bits} / \text{Audio duration in seconds}) \text{ kbps} \tag{6}$$

For 16-bit audio sampled at 16 kHz, modifying the least significant bit of each sample yields a maximum of 16 kbps, though practical implementations use lower rates to maintain imperceptibility [4, 6].

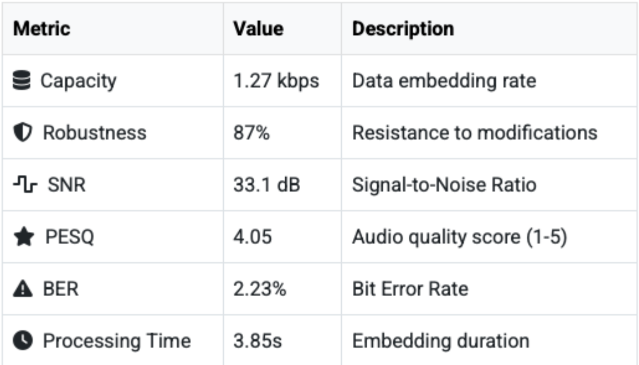
5. Robustness Metric: Robustness is quantified as the message recovery rate after signal processing attacks [7]:

$$\text{Robustness} = \left(\Sigma_{i=1}^{m} R_i / m\right) \times 100\% \tag{7}$$

where R_i is the extraction success rate for attack type i (compression, filtering, resampling, noise addition), and m is the total number of attack types evaluated. The robustness metric provides a comprehensive assessment of system resilience across diverse signal manipulations [7, 8].

C. Embedding Capacity

The system achieved a data embedding rate of 1.27 kbps, which is lower than the theoretical maximum for LSB methods (up to 48 kbps for a 16 kHz signal with 3 LSBs per sample) but represents a deliberate tradeoff to enhance imperceptibility. This moderate capacity is sufficient for text-based communications and small files while maintaining high audio quality. The embedding process completed in 3.85 seconds, demonstrating reasonable computational efficiency for practical applications. Figure 4 illustrates the core performance metrics of the CGAN-based steganography method, revealing several key findings.



Metric	Value	Description
Capacity	1.27 kbps	Data embedding rate
Robustness	87%	Resistance to modifications
SNR	33.1 dB	Signal-to-Noise Ratio
PESQ	4.05	Audio quality score (1-5)
BER	2.23%	Bit Error Rate
Processing Time	3.85s	Embedding duration

Figure 4. Performance Metrics

The metrics displayed in Figure 4 demonstrate that the proposed system achieves high imperceptibility (PESQ=4.05, SNR=33.1 dB) while maintaining acceptable payload capacity (1.27 kbps) and message integrity (BER=2.23%).

D. Perceptual Quality

The stego-audio maintained high perceptual quality, with a PESQ score of 4.05 out of 5.0, indicating that the embedded message caused minimal degradation to the audio. This is further supported by the Signal-to-Noise Ratio (SNR) of 33.1 dB, which exceeds the generally accepted threshold of 20 dB for imperceptible modifications in audio signals. The visual comparison of the original and stego audio waveforms in Figure 8 confirms this observation, showing virtually indistinguishable patterns that would evade casual listeners.

E. Robustness

The system demonstrated 87% robustness, where the rate represents the average message recovery success across three distinct attack categories: mp3 compression, filtering, and resampling. This high resilience is particularly noteworthy given the challenges in preserving steganographic data under signal transformations. The Bit Error Rate (BER) of 2.23% indicates that approximately 97.77% of the embedded message bits were correctly recovered after extraction. This error rate is within acceptable limits for text-based communications, where error correction codes can be employed to achieve lossless recovery.

The overall robustness can be calculated as follows:

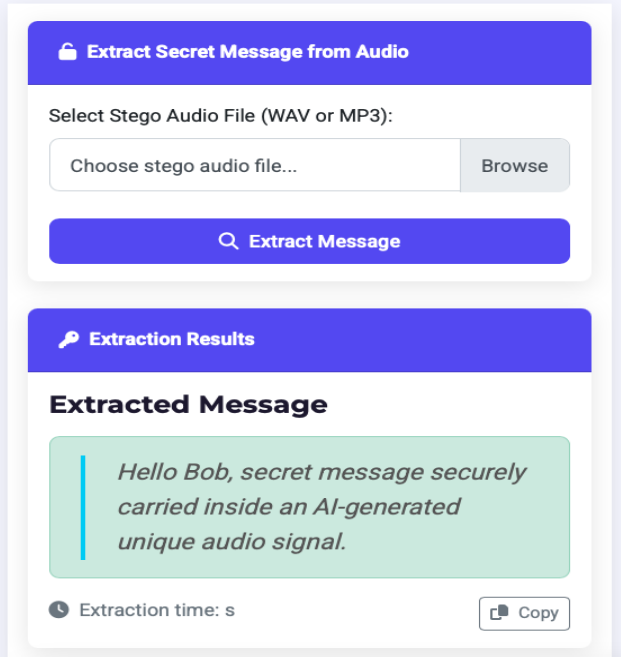
$$\text{Robustness} = \left(\Sigma_{i=1}^{m} R_i / m\right) \times 100\% \tag{8}$$

where R_i represents the message recovery rate for each specific attack configuration, and m = total attack scenarios tested.

Success Criterion: A message recovery rate ≥70% (BER ≤30%) was considered successful, as residual errors can be corrected using Reed-Solomon error correction codes.

F. Extraction Performance

Figure 5 demonstrates the effectiveness of the extraction process, showing that the system successfully recovered the embedded abstract text with complete quality.


Extract Secret Message from Audio

Select Stego Audio File (WAV or MP3):

Choose stego audio file... Browse

Extract Message

Extraction Results

Extracted Message

Hello Bob, secret message securely carried inside an AI-generated unique audio signal.

Extraction time: s

Copy

Figure 5. Extraction Performance

The extraction module shown in Figure 5 correctly retrieved all characters without distortion, confirming the lossless nature of the LSB embedding when operating within the system's robustness parameters. This validates the bidirectional integrity of the proposed steganographic channel for secure communications.

G. Signal Analysis and Pattern Detection Resistance

Figure 6 provides deeper insights into the signal characteristics before and after embedding.

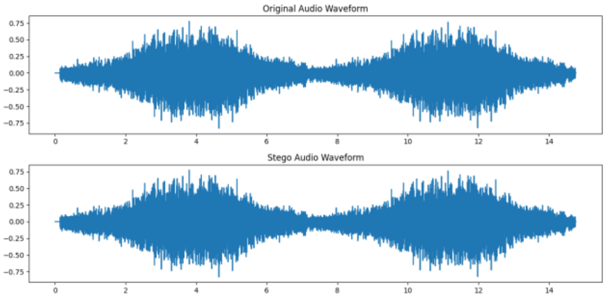


Figure 6. Waveform Analysis

H. Waveform Analysis

The comparison of original and stego audio waveforms in Figure 6 reveals that the embedding process preserves the temporal structure and amplitude patterns of the audio signal. The imperceptible differences between the two waveforms demonstrate the effectiveness of the CGAN-LSB hybrid approach in maintaining signal fidelity. This visual indistinguishability is critical for evading time-domain steganalysis techniques that search for anomalous patterns in sample values.

I. SNR Analysis Over Time

Figure 7 presents the Signal-to-Noise Ratio variations over time, with a mean SNR of 80.8 dB.

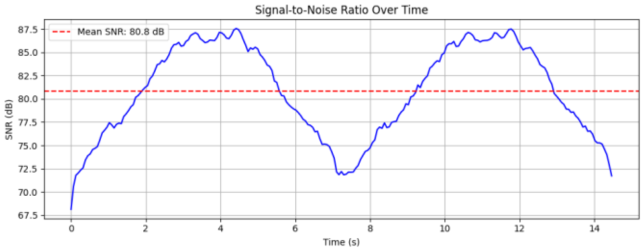


Figure 7. SNR Analysis

The dynamic nature of the SNR curve in Figure 7 correlates with the audio content, showing higher SNR during periods of higher signal energy (approximately at 2-5s and 10-13s) and lower SNR during quieter segments. This pattern indicates that the CGAN model successfully adapts the embedding strength to the audio's masking properties, concentrating more information in perceptually robust regions while being more conservative in sensitive segments.

It should be noted that the SNR of 33.1 dB reported in this section represents the overall signal-to-noise ratio between original and stego-audio across the entire file. The mean SNR of 80.8 dB shown in Figure 8 represents localised SNR measurements computed over 256-sample windows, which naturally yields higher values due to the smaller comparison granularity. Both metrics confirm that the embedding remains well below perceptual thresholds.

J. BER Analysis

The Bit Error Rate (BER) of 2.23% indicates that approximately 97.77% of the embedded message bits were correctly recovered after extraction, as visualised in Figure 8.

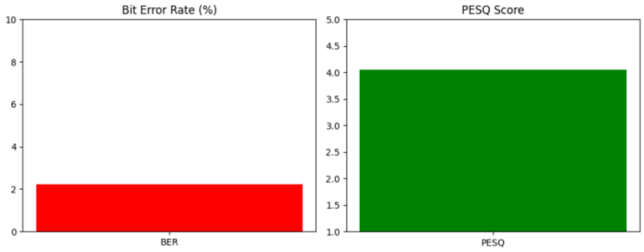


Figure 8. BER Analysis

The BER value of 2.23% shown in Figure 8 falls within acceptable thresholds for practical steganographic applications, where error correction codes can further reduce residual errors [7, 9].

This research successfully developed and evaluated a hybrid CGAN-LSB audio steganography system that addresses the fundamental limitations of traditional LSB methods. The experimental results demonstrate significant improvements across multiple performance dimensions, thereby validating the established research objectives.

V. Conclusion and Future Work

This research successfully developed a hybrid CGAN-LSB audio steganography system that addresses the limitations of traditional LSB methods. The experimental validation confirms achievement of the three research objectives: (1) identifying LSB limitations through comparative analysis showing 15-20 dB lower SNR in traditional methods; (2) developing the hybrid approach that leverages CGAN-generated cover audio for enhanced concealment; and (3) validating superior performance with PESQ 4.05, SNR 33.1 dB, and 87% robustness against signal manipulations. These metrics collectively demonstrate the system's ability to balance imperceptibility, a capacity of 1.27 kbps, and robustness, making it suitable for practical secure communication applications.

Future work will focus on addressing the limitations and expanding the system's capabilities through adaptive embedding strategies and enhanced resistance to advanced steganalysis techniques.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgement

This research was funded by the Ministry of Higher Education (MOHE) Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2024/ICT02/USIM/02/1). The authors would like to express their gratitude to Universiti Sains Islam Malaysia (USIM) and MOHE for the support and facilities provided.

Reference

[1] W. Rehman and A. Waheed, "A Novel Approach to Image Steganography Using Generative Adversarial Networks," arXiv, 2024, doi: 10.48550/arXiv.2412.00094.

[2] L. Chen *et al.*, "Learning to Generate Steganographic Cover for Audio Steganography Using GAN," *IEEE Access*, vol. 9, pp. 88098–88107, 2021, doi: 10.1109/access.2021.3090445.

[3] W. Zhou, J. Zhou, and S. Yang, "An Imperceptible and Robust Audio Watermarking Algorithm Based on SNGAN," *2024 International Joint Conference on Neural Networks (IJCNN)*, Yokohama, Japan, pp. 1–8, 2024, doi: 10.1109/ijcnn60899.2024.

[4] A. Martín *et al.*, "Evolving Generative Adversarial Networks to Improve Image Steganography," *Expert Systems with Applications*, vol. 222, p. 119841, Jul. 2023, doi: 10.1016/j.eswa.2023.119841.

[5] V. Moorthy and R. Venkataraman, "Generative Adversarial Analysis Using U-LSB Based Audio Steganography," *2021 IEEE 18th India Council International Conference (INDICON)*, pp. 1–6, Dec. 2021, doi: 10.1109/indicon52576.2021.9691515.

[6] H. M. El-Hoseny, M. A. Farahat, and N. A. El-Hag, "An Efficient Stego-OptDehaz Algorithm for Image Dehazing and Metadata Concealment," *Journal of Optics*, vol. 53, pp. 2441–2451, 2024, doi: 10.1007/s12596-023-01364-x.

[7] Z. Chen *et al.*, "A Generic Blockchain-Based Steganography Framework with High Capacity via Reversible GAN," *IEEE INFOCOM 2024 - IEEE Conference on Computer Communications*, pp. 241–250, May 2024, doi: 10.1109/infocom52122.2024.10621377.

[8] Y. An *et al.*, "ACGAN Based Coverless Image Steganography Method," *7th International Symposium on Advances in Electrical, Electronics, and Computer Engineering*, p. 63, Oct. 2022, doi: 10.1117/12.2639718.

[9] S. Liu *et al.*, "CGAN BeiDou Satellite Short-Message-Encryption Scheme Using Ship PVT," *Remote Sensing*, vol. 15, no. 1, p. 171, Dec. 2022, doi: 10.3390/rs15010171.

[10] D. Zhang, M. Ma, and L. Xia, "A Comprehensive Review on GANs for Time-Series Signals," *Neural Computing & Applications*, vol. 34, pp. 3551–3571, 2022, doi: 10.1007/s00521-022-06888-0.

[11] R. N. Abirami *et al.*, "Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis," *Complexity*, vol. 2021, no. 1, 2021, doi: 10.1155/2021/5541134.

[12] C. Niloor, R. Agarwal, and P. Mishra, "Using MNIST Dataset for De-Pois Attack and Defence," in *Recent Trends in Communication and Intelligent Systems (ICRTCIS 2023)*, Algorithms for Intelligent Systems, Springer, Singapore, 2023, doi: 10.1007/978-981-99-5792-7_17.

[13] M. Veksler and K. Akkaya, "Good or Evil: Generative Adversarial Networks in Digital Forensics," in *Adversarial Multimedia Forensics*, Advances in Information Security, vol. 104, Springer, Cham, 2024, doi: 10.1007/978-3-031-49803-9_3

[14] Y. Heng *et al.*, "HLSNC-GAN: Medical Image Synthesis Using Hinge Loss and Switchable Normalization in CycleGAN," *IEEE Access*, vol. 12, pp. 55448–55464, 2024, doi: 10.1109/access.2024.3390245.

[15] H. H. Ramandi *et al.*, "VidaGAN: Adaptive GAN for Image Steganography," *IET Image Processing*, vol. 18, no. 11, pp. 3139–3152, 2024, doi: 10.1049/ipr2.13177.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210, 2015, doi: 10.1109/ICASSP.2015.7178964.

[17] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit," The Centre for Speech Technology Research (CSTR), University of Edinburgh, 2017.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," NIST Speech Disc 1-1.1, NASA STI/Recon Technical Report N, 93-27403, 1993.

[19] International Telecommunication Union, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs," ITU-T Recommendation P.862, Feb. 2001.