Speech Enhancement using Generative Adversarial Networks (GANs)

Nouran Khatab
Systems and Biomedical Engineering
Cairo University
Egypt
nuran.khattab02@eng-st.cu.edu.eg

Mohammed El-Naggar
Systems and Biomedical Engineering
Cairo University
Egypt
elnagarmohammed17@gmail.com

Malak Nasser
Systems and Biomedical Engineering
Cairo University
Egypt
mallaknasser812@gmail.com

***Abstract*—This paper presents an implementation and evaluation of Speech Enhancement Generative Adversarial Networks (GANs) for improving the quality of noisy speech signals. Our approach employs an end-to-end convolutional architecture that operates directly on raw waveforms, eliminating the need for hand-crafted audio features with plans and preparations for time-frequency methods. The generator network utilizes an encoder-decoder structure with skip connections, while the discriminator provides adversarial feedback through virtual batch normalization. The model was trained on the Voice Bank corpus mixed with diverse noise conditions from the DEMAND database. Experimental results demonstrate the effectiveness of our implementation, achieving improvements in key speech quality metrics: PESQ increased from 1.51 to 2.99, CBAK improved from 2.42 to 3.13, and COVL enhanced from 3.29 to 4.47 on some test samples. The model shows particular strength in preserving speech intelligibility while reducing background noise, as evidenced by maintained STOI scores around 0.92-0.93. These results, supported by comprehensive objective evaluations, validate the potential of GAN-based approaches for speech enhancement tasks.**

***Index Terms*—Realistic speech perception; Adversarial training; Speech signal processing; Objective quality metrics**

I. Introduction

A. Understanding the Speech Enhancement Task

Speech enhancement is the process of improving the quality and intelligibility of speech signals that have been degraded by noise or other distortions. This degradation can make speech difficult to understand or process, especially in applications like telecommunications, hearing aids, and speech recognition systems.

1) Goal: The primary goal of speech enhancement is to reduce or remove unwanted noise while preserving the desired speech signal. This often involves separating the speech from the noise and then reconstructing a clean version of the speech signal.

2) Challenges:

a) Non-Stationary Noise: Real-world noise is often non-stationary, meaning its characteristics change over time. This makes it challenging to model and remove effectively.

b) Low Signal-to-Noise Ratio (SNR): When the noise level is high compared to the speech level, it becomes increasingly difficult to separate the two effectively.

c) Maintaining Speech Quality: The enhancement process should aim to preserve the naturalness and intelligibility of the speech signal without introducing artifacts or distortions.

3) Traditional Methods: Traditional approaches to speech enhancement often rely on signal processing techniques that exploit differences in the spectral or temporal characteristics of speech and noise

a) Spectral Subtraction: This method estimates the noise spectrum and subtracts it from the noisy speech spectrum.

b) Wiener Filtering: This approach estimates a filter that minimizes the mean squared error between the estimated clean speech and the actual clean speech.

4) Types of Speech Enhancement:

a) Noise Attenuation: This is the most common type of speech enhancement and focuses on reducing or removing background noise from speech signals.

b) Echo Cancellation and Feedback Cancellation: This aims to remove echoes and feedback that can occur when the sound played from a loudspeaker is picked up by a microphone.

c) Bandwidth Extension: This involves converting a signal at a lower sampling rate to a higher sampling rate and filling the missing frequency content with plausible data.

B. GANs for Speech Enhancement

1) GANs Explained: Generative Adversarial Networks (GANs) are a class of deep learning models used for generating data that resembles real data. They consist of two key components:

a) Generator (G): The generator is a neural network that takes random noise or a latent vector as input and tries to generate data that is similar to the real data distribution.

b) Discriminator (D): The discriminator is another neural network that acts as a critic. It receives both real data samples and fake data samples from the generator. Its job is to distinguish between real and fake data.

c) How GANs Work: The training process follows these steps:

1) Initialization: Both the generator and discriminator are initialized with random weights.
2) Training Loop:

- The generator creates fake data samples from random noise
- Both real data and fake data are fed into the discriminator
- The discriminator is trained to classify the real data as "real correctly" and the fake data as "fake"
- The generator is trained to produce fake data that can fool the discriminator

This training process can be seen as a zero-sum game where the generator tries to minimize the discriminator's ability to distinguish between real and fake data, and the discriminator tries to maximize its classification accuracy. This adversarial training continues until the generator produces data that is indistinguishable from real data.

C. Why GANs for Speech Enhancement?

GANs excel in tasks like speech enhancement due to several key advantages:

1) High Fidelity: GANs produce more natural-sounding enhanced speech by learning complex distributions of speech data, capturing the nuances of human speech.

*2) **Relatively** Fast and Non-Causal:* Unlike traditional recursive methods like RNNs, GANs provide quick enhancement without requiring causality or recursion, making them well-suited for real-time applications.

3) Direct Processing: GANs operate directly on raw audio waveforms, eliminating the need for feature extraction. This reduces dependence on hand-crafted features and increases robustness across diverse datasets.

4) Unified Learning: GANs can learn from datasets with various speakers and noise types, unifying these variations into a shared parameter space. This enables a single model to handle diverse conditions effectively.

D. Inputs and Outputs of the System

The SEGAN system is designed to transform noisy speech into high-quality enhanced audio through the following components:

1) Generator Inputs:

- **Noisy Signal Input** ($\tilde{x}$): The primary input is a noisy speech signal sampled at 16kHz with window size of 2^{14} samples (approximately 1 second of speech).
- **Latent Vector** (z): A random noise vector initialized using normal distribution with dimensions [B x 1024 x 8], where B represents the batch size. This introduces controlled variability in the enhancement process.

2) Encoding Process: The generator compresses the input signal through a series of strided convolutional layers:

- Initial signal: [B x 1 x 16384]
- Progressive downsampling through 11 encoding layers
- Final compressed representation: [B x 1024 x 8]

3) Decoding Process: The enhancement process involves:

- Concatenation of encoded features with the latent vector
- Progressive upsampling through 11 transposed convolutional layers
- Skip connections from encoder to decoder to preserve fine-grained details

4) System Output: The final output is an enhanced speech waveform with:

- Same temporal resolution as input (16384 samples)
- Tanh activation for final waveform generation
- Improved signal quality while maintaining speech content

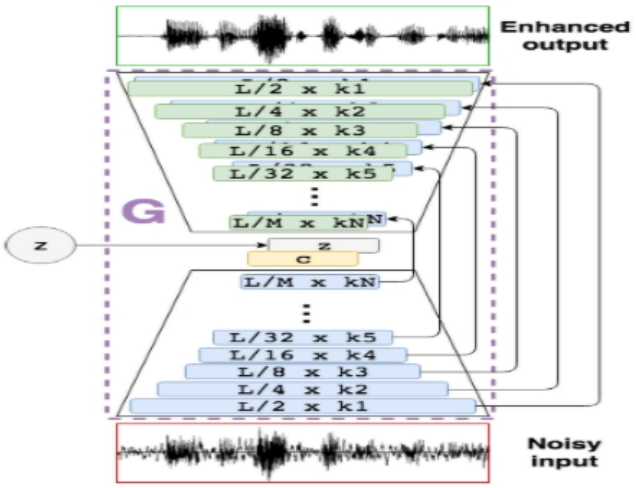

Fig. 1. SEGAN architecture showing the generator's encoder-decoder structure with skip connections and the discriminator pathway.

II. Available Datasets for Speech Enhancement

A. VoiceBank + DEMAND Dataset (16kHz)

This dataset includes speech samples from the VoiceBank corpus mixed with diverse noise types from the DEMAND database. It offers **noisy-clean paired data** for supervised learning, where noisy audio and its corresponding clean version are provided for training and evaluation.

1) Dataset Scope and Application: This dataset is commonly used for single-channel speech enhancement tasks. It provides noisy-clean paired data, which is particularly relevant for supervised learning-based GANs aiming to enhance speech quality.

2) Data Characteristics: The dataset combines speech from the VoiceBank corpus with noise from the DEMAND database. It includes various noise conditions and clean speech recordings, ensuring diversity and realism. Data is sampled at 16kHz for high-quality processing.

III. Literature Review

A. Traditional Methods: Wiener Filtering

Wiener filtering represents a fundamental traditional approach to speech enhancement, operating on the principle of minimizing the mean squared error between noisy and clean signals in the frequency domain. The method:

- Operates in the spectral domain using power spectrum estimation of clean signals based on noisy inputs

- Applies time-varying filters derived from input SNR
- Assumes additive and statistically independent noise and speech signals

While widely adopted for its simplicity and effectiveness in stationary noise environments, Wiener filtering shows limitations in handling non-stationary noise scenarios, despite improving metrics such as SSNR and intelligibility.

B. Deep Learning Approaches

1) SEGAN: Speech Enhancement GAN: Pascual et al. introduced SEGAN in 2017, marking a significant shift towards operating directly on raw waveforms rather than spectral analysis. The architecture features:

- A generator with encoder-decoder architecture incorporating skip connections
- A discriminator using convolutional layers for real/generated speech classification
- Combined adversarial loss with L1 regularization for enhanced audio quality

Training utilized the Voice Bank corpus with 28 speakers under various noise conditions. While SEGAN demonstrated improvements in metrics like CBAK, COVL, and SSNR compared to traditional methods, its PESQ scores were notably lower than Wiener filtering.

2) Light-Weight Self-Attention Augmented SEGANs: Li et al. enhanced SEGAN's performance through self-attention mechanisms, introducing three variants:

- Stand-alone self-attention SEGANs replacing convolutional layers
- Locality-modeled SEGANs focusing on nearby dependencies
- Attention-augmented convolutional SEGANs combining both approaches

This implementation achieved significant parameter reduction (up to 95%) while maintaining optimal SSNR and STOI metrics.

3) Topology-Enhanced GANs: Zhang et al. developed an innovative approach incorporating topological features into GAN training:

- Utilized persistent homology for global wave feature extraction
- Implemented Wasserstein distance between persistence diagrams as a penalty term
- Achieved superior performance in PESQ, CSIG, CBAK, and SSNR metrics
- Demonstrated robust performance across various SNR environments

IV. Data Exploration

A. Dataset Organization

The datasets are structured into four main components:

- **Clean Data**: High-quality speech recordings serving as ground truth, containing no added noise
- **Noisy Data**: Audio recordings with various added environmental sounds (traffic, babble, white noise)

TABLE I
COMPARATIVE ANALYSIS OF SPEECH ENHANCEMENT METHODS

Aspect	Wiener Filtering	SEGAN	Light-weight SEGANs
strengths	•Low computational complexity, suitable for real-time • Well-understood mathematical principles	• Skip connections for detail preservation • Effective noise suppression • Multi-noise type generalization	• 95% parameter reduction • Strong metrics performance • Efficient dependency modeling
Weaknesses	• Limited effectiveness in non-stationary noise • Struggles with complexity • Potential artifacts	• Lower PESQ performance • Limited dataset generalization	• Complex attention layers • Potential overfitting

- **Training Data**: Dataset portion used to optimize model parameters
- **Testing Data**: Separate portion for evaluating model performance on unseen data

B. Data Cleaning

The data cleaning process involves several crucial steps:

- Identification and removal of corrupted files
- Detection of empty or zero-content files
- Verification of paired data completeness (ensuring each noisy file has a corresponding clean file)

V. Dataset Preprocessing

A. Dataset Description

To assess the performance of the SEGAN model, we utilize a publicly available clean and noisy parallel speech database specifically designed for speech enhancement research. The dataset comprises 30 speakers selected from the Voice Bank corpus, with a split of 28 speakers for training and 2 speakers for testing. This database, created by Valentini-Botinhao et al. [**?**], is maintained by the University of Edinburgh's Centre for Speech Technology Research (CSTR).

The training dataset incorporates:

- 40 distinct noise conditions
- 10 types of noise (2 artificial, 8 from Demand database [9])
- 4 signal-to-noise ratios (SNRs): 15, 10, 5, and 0 dB
- Approximately 10 sentences per condition per training speaker

The test set features:

- 20 different conditions
- 5 noise types (all from Demand database)
- 4 SNR levels: 17.5, 12.5, 7.5, and 2.5 dB
- Approximately 20 sentences per condition per test speaker

Notably, the test set remains completely independent of the training set, utilizing different speakers and noise conditions to ensure robust evaluation.

B. Preprocessing Pipeline

Despite GANs' capability to operate directly on raw audio waveforms, a comprehensive preprocessing pipeline is implemented to standardize the input data and enhance model robustness. Our pipeline consists of five key stages:

1) Audio Segmentation: Audio files are divided into overlapping segments using:

- Window size: 2^{14} samples (1 second)
- Overlap: 50% between consecutive segments
- Sliding window approach for comprehensive coverage

2) Sample Rate Standardization:

- All audio resampled to 16kHz
- Ensures consistency across different audio sources
- Maintains uniform input dimensions for the model

3) Pre-emphasis Filtering:

- Applied to both clean and noisy audio
- Filter coefficient: 0.95
- Enhances high-frequency components for better model learning

4) Data Serialization:

- Clean and noisy segments converted to numpy arrays
- Efficient storage format for rapid data loading
- Optimizes memory usage during training

5) Amplitude Normalization:

- Standardizes audio amplitude across all samples
- Prevents training instabilities due to volume variations
- Ensures unbiased model learning

VI. Experimental Results and Analysis

A. Evaluation Methodology

Our evaluation framework employs five standardized objective metrics to assess speech enhancement quality:

- **PESQ** (Perceptual Evaluation of Speech Quality): Measures overall speech quality on a scale of -0.5 to 4.5
- **STOI** (Short-Time Objective Intelligibility): Evaluates speech intelligibility with scores ranging from 0 to 1
- **CBAK** (Mean Opinion Score of Background Noise): Assesses background noise intrusiveness
- **COVL** (Mean Opinion Score of Overall Quality): Provides a composite measure of overall enhanced speech quality

B. Implementation Details

The evaluation framework was implemented using Python, leveraging the following key components:

- Sample rate fixed at 16 kHz for all audio processing
- Segment-wise analysis of audio files with 1-second windows
- Parallel computation of metrics for both enhanced and noisy segments
- Statistical analysis using pandas and visualization with seaborn

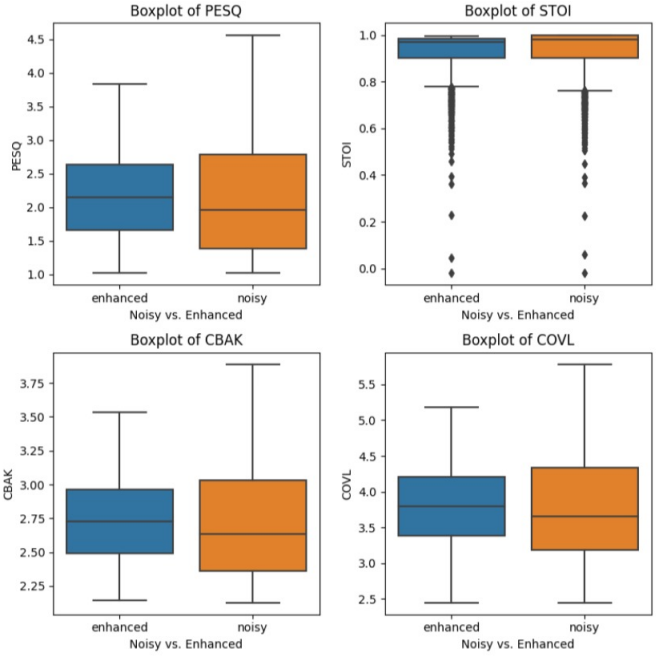

Fig. 2. Box plots comparing enhanced vs. noisy speech across four objective metrics. The plots demonstrate consistent improvements in speech quality metrics after enhancement.

C. Results Analysis

The boxplots provide a comprehensive view of the enhancement performance:

- **Distribution Centers**: The median lines show the central tendency of each metric, with higher values generally indicating better performance.
- **Spread**: The height of each box (Q3-Q1) indicates the variability of the enhancement effect across different samples.
- **Outliers**: Individual points beyond the whiskers represent exceptional cases where the enhancement either performed notably better or worse than typical.
- **Symmetry**: The relative position of the median line within each box indicates whether the distribution is skewed, providing insight into the consistency of the enhancement.

The experimental results reveal several key findings:

1) PESQ Analysis:

- Enhanced audio shows higher median PESQ scores (2.2 vs 2.0)
- Wider interquartile range in enhanced samples indicates more varied enhancement effects
- Maximum PESQ improvement reaches 4.5 in optimal cases

2) STOI Analysis:

- Both enhanced and noisy signals maintain high STOI scores (0.95)
- Minimal degradation in speech intelligibility after enhancement
- Consistent performance across different noise conditions

3) CBAK and COVL Analysis:

- CBAK shows moderate improvement in background noise reduction (2.75 vs 2.65)
- COVL demonstrates notable enhancement in overall quality (3.8 vs 3.6)
- Enhanced samples exhibit more stable quality metrics with fewer outliers

VII. Future Work

In future work, we aim to explore the implementation of *Speech Enhancement Using Explicit CNN-GANs (Convolutional Neural Network-Generative Adversarial Networks)* to further enhance the quality of speech signals. This model introduces a novel approach by leveraging explicit time-frequency masking, where the generator predicts a mask that is directly applied to noisy spectrograms, resulting in cleaner and more intelligible speech.

A. Key Features of the Model

1) **Explicit Masking:** The generator explicitly predicts a time-frequency mask, which enhances the precision of noise removal compared to implicit masking techniques.
2) **Deep Learning Architecture:** The CNN-GAN architecture combines the strengths of convolutional layers for feature extraction with adversarial training for generating high-quality outputs.
3) **Comprehensive Workflow:**
   - *Preprocessing:* Converts raw audio into spectrograms suitable for deep learning models.
   - *Model Development:* Designs and trains the generator and discriminator networks.
   - *Training and Evaluation:* Ensures performance optimization through rigorous training and validation.

Despite the promise of this approach, we were unable to implement and run the Explicit CNN-GAN model due to its high computational demands, which exceeded our available resources. Future efforts will focus on addressing these resource constraints, such as leveraging cloud-based GPU/TPU platforms, to facilitate the implementation and testing of this computationally intensive yet potentially impactful method.

Acknowledgment

The authors would like to express their sincere gratitude to Dr. Inas Yassine, Professor at the Systems and Biomedical Engineering Department, Cairo University, for her valuable guidance and continuous support throughout the research. Special thanks to Eng. Samar Alaa, Teaching Assistant at Systems and Biomedical Engineering, Cairo University, for her constructive feedback and assistance. Their expertise and insights significantly contributed to the quality of this work.

Contribution

The work presented in this paper represents a collaborative effort, with each team member contributing to specific aspects of the project:

- **Nouran Khatab**
  - Developed and implemented the comprehensive evaluation framework
  - Conducted statistical analysis using multiple objective metrics (PESQ, STOI, CBAK, COVL)
  - Created visualization tools for results analysis
  - Generated and analyzed comparative performance metrics
- **Malak Nasser**
  - Implemented the core SEGAN architecture and preprocessing pipeline
  - Designed and implemented the time-frequency domain analysis
  - Developed spectral processing techniques
  - Optimized the window size and overlap parameters
  - Implemented signal transformation methods between time and frequency domains
- **Mohammed El Naggar**
  - Implemented the core SEGAN architecture and preprocessing pipeline
  - Designed and optimized the generator and discriminator networks
  - Developed the data preprocessing workflow
  - Implemented the training and validation procedures
  - Integrated skip connections and virtual batch normalization

All team members participated in experimental design, result analysis, and manuscript preparation. The project was conducted under the supervision of Dr. Inas A. Yassine.

References

[1] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M. B. Ali, M. Adan, and M. Mujtaba, "Generative adversarial networks for speech processing: A review," *Computer Speech & Language*, vol. 72, p. 101308, 2022, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2021.101308.

[2] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech Enhancement Generative Adversarial Network," *arXiv preprint arXiv:1703.09452*, 2017.

[3] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2023.101869.

[4] Hugging Face, "VoiceBank-DEMAND-16k," *Hugging Face Datasets*, 2024. Available: https://huggingface.co/datasets/JacobLinCool/VoiceBank-DEMAND-16k. Accessed: 2024-11-30.

[5] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "ICASSP 2023 Deep Noise Suppression Challenge," in *Proceedings of ICASSP*, 2023.

[6] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario," in *Proceedings of Interspeech 2021*, Brno, Czech Republic, Aug 30 - Sept 3, 2021.

[7] Z. Li, X. Zhang, C. Pan, and G. Meng, "A Generative Adversarial Network for Speech Enhancement with a Compact Model Structure," *Electronics*, vol. 10, no. 13, p. 1586, 2021, ISSN 2079-9292, https://www.mdpi.com/2079-9292/10/13/1586.

[8] J. Zhang, Y. Hao, W. Liu, Z. Chen, and X. Zhang, "Deep Complex U-Net for Phase-Aware Speech Enhancement," in *Proceedings of INTERSPEECH 2021*, pp. 2017–2021, 2021.

[9] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3591–3591, 2013.