Volume 65, Number 1, 2025

**ACTA TECHNICA NAPOCENSIS**
*Electronics and Telecommunications*

# AUDIO DENOISING USING U-NET ARCHITECTURE

L.-Daniel JIMON, Mircea-F. VAIDA
*Technical University of Cluj-Napoca, 26-28 Baritiu str. Cluj-Napoca, Romania*
*jimon.mi.lucian@student.utcluj.ro, mircea.vaida@com.utcluj.ro*

**Abstract: Audio denoising is a pivotal task in audio signal processing. This paper presents a machine learning approach using a U-Net architecture to denoise musical audio signals affected by four distinct noise types: white noise, urban noise, reverberation, and noise cancellation artifacts. The model was evaluated on datasets derived from IRMAS and UrbanSound8K. Objective and subjective evaluation metrics were used, which show the model's effectiveness in filtering white and urban noise. However, performance on reverberation and noise cancellation artifacts is limited, indicating areas for future architectural and methodological improvements.**
***Keywords:*** *audio denoising, machine learning, U-Net, deep learning, signal processing.*

## I. INTRODUCTION

Noise reduction plays a critical role in audio signal processing, significantly enhancing the quality and intelligibility of signals in applications such as speech signal processing [1], [2], [3], music production and restoration [4], and biology [5]. Traditional denoising methods, including spectral subtraction and Wiener filtering, have limitations when dealing with complex noise patterns, even though are considered fundamental tools in denoising tasks [6], [7]. With advancements in deep learning, convolutional neural networks (CNNs) have been successfully applied to denoising tasks, particularly in the form of encoder-decoder architectures such as U-Net [8].

This paper investigates the effectiveness of a rather simple U-Net-based machine learning model for denoising musical audio signals, using spectrogram transformations for input representation. The model was trained separately with different noise types, thus demonstrating its strengths and weaknesses in real-world scenarios (for example, music restoration, homemade music recording etc.).

This paper is organized as follows. Section II reviews the state-of-the-art in audio denoising. Section III describes the datasets and noise types used for training and evaluation. Section IV details the U-Net model architecture and its mathematical foundations. Section V outlines the training procedure, while Section VI presents the evaluation methodology and discusses the quantitative and qualitative results. Finally, Section VII concludes the paper and suggests directions for future work.

## II. STATE-OF-THE-ART

Audio denoising, even though a thoroughly researched subject in signal processing, has been primarily discussed in speech enhancement circumstances, therefore most of the breakthroughs and ideas come from papers discussing voice signal denoising. The most used methods are mainly LSTM-based systems [2], auto-encoders [9] and CNN-related methods [1], [10]. More recently, new approaches has been investigated, based on state-of-the-art technologies, which include transformer-based architectures like DPT-FSNet [11] and generative adversarial networks (GANs) such as CMGAN [12], which improve time-frequency processing capabilities.

As mentioned above, most studies focus on speech signal denoising, with limited emphasis on musical audio. This paper adapts the U-Net model to denoise musical signals, addressing a research gap in the application of CNNs for this purpose.

## III. DATASETS AND NOISE TYPES

Selecting an adequate dataset is fundamental in designing the denoising model, since the process depends on how well the dataset's samples are illustrating the real-life scenarios presented in the first section of the current paper, and how the chosen noise types interact with the diverse elements from each musical sound used for training. A well-balanced, preferably large enough dataset with audio files having standardized length and sample rate benefits the preprocessing part of the model training.

Additionally, the noise introduced should be realistic and varied, capturing both synthetic and natural disturbances that can be usually found in affected audio signals.

### *A. Clean dataset*

The chosen dataset which was used as 'clean dataset' in the presented study, is the IRMAS dataset [13], which contains 6705 audio files being 2s long and which were extracted from recordings sampled at 44.1kHz.

The samples are all musical audio files and present a great variation of sounds: from single instrument recordings to orchestras or pop-music bands.

### *B. Noise types*

For a comprehensive simulation of real-life sound alterations, the following noise types were implemented and added to the clean musical samples: *white noise, urban noise, reverberation, and noise cancellation artifacts*, each discussed below.

The white noise was manually added to the IRMAS samples, by generation of Gaussian noise and adding it to the clean audio.

To simulate real-world recording conditions, urban noise was sourced from the UrbanSound8K dataset [14]. Ten

*Manuscript received April 17, 2025; revised June 27, 2025*

representative 2-second samples, sampled at 44.1kHz to match the IRMAS dataset, were selected. These samples include sounds such as dog barking, traffic, human speech, and construction noise, and were combined with the clean musical samples.

The reverberation effect, common in rooms with poor acoustics, was simulated using the Pedalboard Python module [15] from Spotify. Specifically, the Reverb effect was applied with the following parameters to create a consistent and significant reverberant field: room_size=0.9, damping=0.9, and wet_level=0.33.

Lastly, noise cancellation effects were simulated by applying partial attenuation and frequency suppression to clean signals. This type of noise commonly appears in low-budget headsets and microphones, whose noise-cancellation systems filter useful sounds as well as noise, from time to time.

In Table 1, a conclusive description of the different noise types can be seen for retrospective purposes.

*Table 1: Used noise types*

| Noise type | Implementation | Purpose |
|---|---|---|
| **White** | Generation Gaussian White noise and adding it to the clean sample | Simulating white noise |
| **Urban noise** | UrbanSound8K Dataset used; clean sample combined with noise sample | Illustrate real life recording situations |
| **Reverberation** | Using Pedalboard module's functionalities | Illustrate low-budget recording circumstances |
| **Noise cancellation** | Adding partial attenuations/frequency suppressions to the clean sample | Simulate use of low-quality microphone |

### C. Training data

The training process used each clean data with its 4 different noise types for the different training run (one for each noise type). The training data consists of spectrogram representations using the Short-Time Fourier Transform (STFT), transformed into NumPy array objects for a lower computational effort during the training.

## IV. MODEL ARCHITECTURE

### A. U-NET

The implemented model is based on the U-Net architecture, a convolutional encoder-decoder network designed for image-to-image tasks, first created for biomedical uses [16].

U-Nets are a type of convolutional neural networks (CNN) primarily designed for image-to-image tasks, such as segmentation. It follows an encoder-decoder structure, where the encoder downsamples the input to capture high-level features, while the decoder upsamples the representation to restore the spatial resolution, this can be seen in Figure 1. A key characteristic of U-Net is its skip connections, which allow information to be transferred directly from corresponding encoder layers to decoder layers [16], [17]. These connections help preserve fine details lost during downsampling, making U-Net effective in applications requiring accurate reconstruction, such as medical image segmentation and audio spectrogram processing. Initially developed for biomedical applications, U-Net has since been adopted in numerous fields ranging from remote sensing to speech enhancement.

When applied to audio denoising, U-Net operates on spectrograms, which are treated as images, allowing the model to use 2D convolutions for feature extraction. By learning the spectral patterns of clean audio, U-Net can effectively suppress unwanted noise while maintaining the integrity of the original signal.

Despite their performance, the U-Net-based models have limitations, particularly when facing long-range dependencies in audio, where transformer-based models or recurrent architectures may provide complementary benefits [11].

### B. Mathematical Model

The U-Net architecture is constructed from several fundamental operations. The input to the proposed model is a 2D spectrogram $\bar{X} \in \mathbb{R}^{FxT}$, where F is the number of frequency bins and T is the number of time frames.

The encoder path consists of repeated blocks of two 2D convolutional layers, each followed by a Rectified Linear Unit (ReLU) activation function, and then a max pooling layer. A 2D convolution is defined as Equation (1) presents below.

$$(X * K)(i,j) = \sum_{m=0}^{i} \sum_{n=0}^{j} X(i-m, j-n) K(m,n) \qquad (1)$$

In the Equation above, $X$ is the input feature map from the previous layer, $K$ is the learnable kernel, and the indices $m$ and $n$ iterate over the dimensions of the kernel. The ReLU activation, $f(x) = \max(0, x)$ introduces non-linearity. The max pooling layer downsamples the feature maps, reducing spatial dimensions while retaining the most prominent features.

The decoder path uses transposed convolutions to upsample the feature maps. These are concatenated with the corresponding feature maps from the encoder path via skip connections. This concatenation is followed by two standard convolutional layers with ReLU activations.

The final layer is a 1x1 convolution that maps the feature channels back to a single channel, producing the final denoised spectrogram $\hat{Y}$.

### C. Network structure

As this paper can be considered a PoC rather than a breakthrough, a rather simple U-Net architecture was chosen, which could be implemented and trained on a not particularly performant machine. The structure can be seen in Figure 1, and it is described below.

*1)* ***Encoder***: 4 downsampling layers, each consisting of a double convolutional block followed by max pooling.

*2)* ***Bottleneck****:* A double convolutional layer to extract deep representations.

*3)* ***Decoder****:* 4 upsampling layers with transposed convolutions and skip connections to retain spatial information.

*4)* ***Final Output****:* A 1-channel spectrogram reconstruction representing the estimated clean audio.

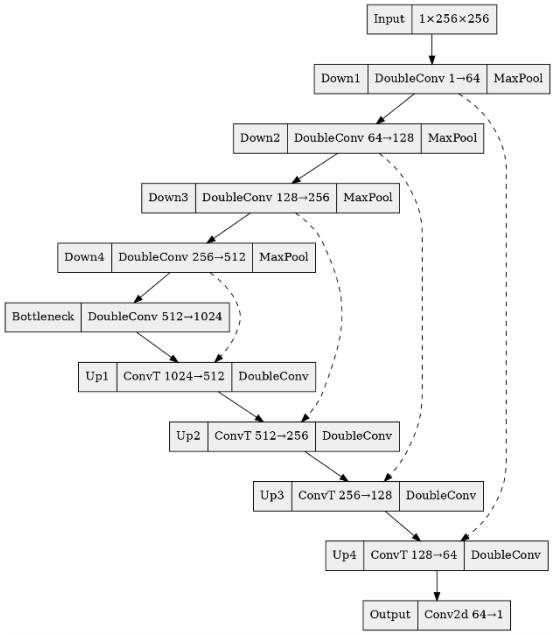


*Figure 1: Implemented U-Net architecture*

## V. TRAINING PROCEDURE

### A. Data Preparation

The configured dataset consisting of clean-noisy pairs was split into 90% training set and 10% validation set. Each audio file was converted into a spectrogram with STFT parameters: FFT size = 255, hop length = 63 (which were determined after a try-and-error iteration) and the resulted dataset was used in batches of 8, which increased training time, but significantly lowered processing power needs.

### B. Training Parameters

The following training configuration was chosen:

*1) Loss function:* Mean Square Error

*2) Optimizer:* Adam (with learning rate = 3e-4)

*3) Epochs:* 20 per noise type

*4) Device:* GPU

As mentioned, four different trained models will result after the training process, one for each noise type. This is allows targeted learning and better generalization, as well as a better understanding of advantages and limitations of U-Net models in denoising applications.

## VI. EVALUATION AND RESULTS

The performance of the trained models was evaluated using both objective, quantitative metrics and subjective, qualitative analysis. Denoised spectrograms were first generated from the test set, then converted back into audio waveforms using the Griffin-Lim algorithm [18] for evaluation.

### A. Evaluation Metrics

To provide a rigorous and standardized assessment, the following objective metrics were employed:

- ***Signal-to-Noise Ration (SNR)***: This metric measures the ratio of the power of the clean signal to the power of the noise. It is expressed in decibels (dB). A higher SNR value indicates better denoising performance. We report both the input SNR (clean vs. noisy audio) and output SNR (clean vs. denoised audio).
- ***Perceptual Evaluation of Speech Quality (PESQ)***: An objective metric standardized by ITU-T P.862 [19] that predicts the subjective quality of a speech signal. It compares the clean reference signal to the degraded (denoised) signal and produces a score from -0.5 to 4.5, where higher scores indicate better quality.

For subjective evaluation, a Mean Opinion Score (MOS) test was conducted, as recommended by ITU-T P.800 [20]. Ten listeners were asked to rate the quality of the denoised audio on a 5-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent).

### B. Quantitative Results

To assess performance, the models were tested on noisy audio files generated with a target input SNR of 5 dB. Due to the varying characteristics of each noise type, the actual measured input SNR differs. Table 2 presents the average objective metrics for each noise type based on these test sets.

*Table 2: Objective Evaluation Results for Noisy Audio Generated at a Target Input SNR of 5 dB*

| Noise type | Average Input SNR (dB) | Average Output SNR (dB) | Average PESQ |
|---|---|---|---|
| **White** | 8.00 | 2.21 | 2.18 |
| **Urban noise** | 8.00 | 2.94 | 2.11 |
| **Reverberation** | 5.76 | 2.58 | 2.90 |
| **Noise cancellation** | 11.48 | 2.07 | 1.88 |

The objective results in Table 2 provide a clear view regarding the model's performance. For every noise type, the Average Output SNR is significantly lower than the Average Input SNR. This indicates that while the model is altering the signal, it is doing so at the cost of damaging the original musical content, leading to a poorer signal-to-noise ratio in the final output. The low PESQ scores, which fall into the "poor" to "fair" range (on a scale of 1 to 4.5), further emphasize this finding, suggesting that the perceptual quality of the denoised audio is low.

Consistent with findings in other domains such as image and speech processing, the shown results confirm that a

standard U-Net architecture has limitations in high-fidelity musical audio denoising. While capable of handling simpler additive noise, it is less effective against complex, non-additive noise types like reverberation and noise cancellation artifacts. The model appears to be too aggressive, removing useful signal components along with the noise, which highlights the need for more advanced model architectures and perceptually-based loss functions for this specific application.

When compared with results from relevant literature, the proposed model demonstrates the limitations of a simple architecture. For instance, a key innovation in this area is the Wave-U-Net [21], a model that applies the U-Net architecture directly to the time-domain waveform. In the original Wave-U-Net paper, the authors' proposed model achieved a median Signal-to-Distortion Ratio (SDR) of 3.49 dB, outperforming a comparable spectrogram-based U-Net which scored 2.74 dB. This highlights that architectural choices and operating domain (time vs. frequency) have a significant impact on performance.

While direct comparison is challenging due to different evaluation metrics (SNR vs. SDR) and datasets, the results are promising, especially in perceptual aspects, presented in the following subsection of this paper. However, the performance of the Wave-U-Net highlights the potential benefits of time-domain processing, and more advanced architectures like DPT-FSNet [11] show far better performance in complex scenarios, indicating clear directions for future work.

### C. Qualitative Results

The reconstructed audio files were subjectively analyzed by the paper's author, and the following observations were noted:

*a) White & Urban Noise:* Denoising was effective with minimal artifacts. However, the urban denoised audios present a non-uniform filtering, meaning that towards the end of the audio segment, the urban noise seemed to have been slowly ignored.

*b) Reverberation:* The model failed to fully recover clean signals due to phase distortions, and additional white noise was introduced; however, a sort of diminution can be sensed in the reverb's effect.

*c) Noise cancellation:* The model performed poorly, requiring more sophisticated techniques, since the cancellation artifacts remained intact.

After the initial subjective analysis, 10 test subjects were chosen and asked to mark the denoised audio files on a 5-point MOS scale (1-5), where 1 means the result is poor and 5 means the result is excellent. The following average marks were obtained, as shown in Table 3.

*Table 3: Average marks after subjective evaluation*

| Noise type | Average MOS (1-5) |
|---|---|
| **White** | 4.1 |
| **Urban noise** | 3.7 |
| **Reverberation** | 1.9 |
| **Noise cancellation** | 2.0 |

The MOS scores confirm the objective findings. Listeners rated the denoising for white and urban noise as "Good" to "Fair," while the results for reverberation and noise cancellation were rated as "Poor," indicating that the artifacts were highly perceptible and bothering.

### D. Discussion of Limitations

While the present study successfully demonstrates the U-Net's capability for certain denoising tasks, it is important to acknowledge its limitations to provide context for the results and guide future research.

*a) Dataset and Generalization:* The training and testing were performed on 2-second audio clips from the IRMAS and UrbanSound8K datasets. While diverse, these datasets may not fully represent the vast array of musical genres, instrumentation, and real-world recording conditions. The short duration of the clips also limits the model's ability to learn long-term temporal dependencies in music, which may be crucial for understanding more complex noise structures.

*b) Architectural Constraints:* As noted, a standard U-Net architecture was intentionally chosen for this proof-of-concept approach. This architecture, while effective for image-like representations, does not have the capability to model long-range dependencies or handle phase information. The poor performance on reverberation is a direct consequence of this, as reverberation is a complex phenomenon involving both magnitude and phase distortions over time. The model's failure directly shows that simply treating a spectrogram as an image is insufficient for phase-sensitive problems.

*c) Phase Reconstruction Artifacts:* The use of the Griffin-Lim algorithm for waveform reconstruction is another limitation. This algorithm estimates the phase from the magnitude spectrogram and is known to introduce audible artifacts, especially after a low number of iterations. Therefore, the final audio quality is influenced not only by the U-Net model's denoising performance but also by the imperfections of the reconstruction algorithm. This could potentially confound the evaluation, as it is difficult to separate artifacts introduced by the model from those introduced by Griffin-Lim.

## VII. CONCLUSION AND FUTURE WORK

This study successfully demonstrated the benefits of using a U-Net-based deep learning model for audio denoising, serving as a proof of concept (PoC) for filtering noise from musical signals. The results, evaluated using standard objective (SNR, PESQ) and subjective (MOS) metrics, indicate that the model performs acceptably on additive noise types like white and urban noise.

However, the model struggled with reverberation and noise cancellation artifacts, which present additional complexities due to phase distortions and spectral modifications. These limitations suggest that a trivial U-Net architecture may not be adequate for handling all types of real-world noise, in special those that introduce non-trivial spectral characteristics.

To improve performance, several future enhancements can be implemented as follows.

*a) Advanced Architectures:* Future work should explore more sophisticated architectures. Integrating residual connections (ResNet blocks ) could allow for the training of much deeper networks without suffering from vanishing gradients or exploding gradients, potentially helping the model to learn more complex features. Furthermore, investigating generative adversarial networks (GANs *[22]*) is a promising direction. A GAN-based model could use a discriminator network to learn a loss function that pushes the denoised audio to be perceptually indistinguishable from clean audio, potentially yielding more natural-sounding results than MSE-based training.

*b) Hybrid Loss Functions:* The use of a simple Mean Squared Error loss function is a major area for improvement. Future research should focus on hybrid loss functions that combine MSE with perceptually-based metrics. For example, a loss function could incorporate an approximation of PESQ or STOI, directly optimizing the network for metrics that better correlate with human hearing and sound quality assessment better than the numerical-only approach tested and presented before, which have resulted in different performances, as Chapter VI described.

*c) Explicit Phase Estimation:* To overcome the limitations of the Griffin-Lim algorithm and better handle the reverberation effect, models that estimate both the magnitude and phase of the clean signal should be investigated. This could mean using a two-branch network where one branch predicts the spectrogram magnitude and the other predicts the phase or a complex-valued network that operates on the complex STFT directly.

*d) Larger and More Realistic Datasets:* To improve generalization, the model should be trained and evaluated on a larger and more varied dataset, including longer audio clips and a wider range of musical styles. Simulating more realistic noise cancellation artifacts and reverberation using varied Room Impulse Responses (RIRs) would also be crucial for enhancing the model's robustness in real-world applications.

In conclusion, the implemented model has shown significant potential in filtering several noise types, and by addressing the mentioned areas, future research can build upon this foundation to develop more effective and adaptive audio denoising solutions.

## VIII. REFERENCES

[1] T. W. Craig Macartney, "Improved Speech Enhancement with the Wave-U-Net," 27 11 2018. [Online]. Available: https://arxiv.org/abs/1811.11307. [Accessed 12 01 2025].

[2] H. E. S. W. E. V. J. L. R. e. a. Felix Weninger, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA), Liberec, Czech Republic, 2015.

[3] G. F. Germain, Q. Chen and V. Koltun, "Speech Denoising with Deep Feature Losses," 06 27 2018. [Online]. Available: https://arxiv.org/abs/1806.10522. [Accessed 12 01 2025].

[4] B. G. M. T. D. R. Yunpeng Li, "Learning to Denoise Historical Music," 05 08 2020. [Online]. Available: https://arxiv.org/abs/2008.02027. [Accessed 12 01 2025].

[5] N. Priyadarshani, S. Marsland, I. Castro and A. Punchihewa, "Birdsong Denoising Using Wavelets," PLoS ONE, vol. 1, pp. 1-11, 2016.

[6] J. Chen, J. Benesty, Y. Huang and S. Doclo, "New Insights Into the Noise Reduction Wiener Filter," IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, vol. 14, no. 4, pp. 1218-1234, 2006.

[7] J. Le Roux and E. Vincent, "Consistent Wiener Filtering for Audio Source Separation," IEEE SIGNAL PROCESSING LETTER, vol. 20, no. 3, pp. 217-220, 2012.

[8] Z. Kang, Z. Huang and C. Lu, "Speech Enhancement Using U-Net with Compressed Sensing," Applied Sciences, vol. 12, no. 9, 2022.

[9] Y. T. e. a. Xugang lu, "Speech enhancement based on deep denoising Auto-Encoder," in Interspeech 2013, 2013.

[10] Vaibhavtalekar, "A Deep Dive into Audio Denoising with TensorFlow(CNN)," Medium, 29 01 2023. [Online]. Available: https://medium.com/@vaibhavtalekar87/a-deep-dive-into-audio-denoising-with-tensorflow-cnn-a996e0c62e16. [Accessed 12 01 2025].

[11] H. C. P. Z. Feng Dang, "DPT-FSNet: Dual-path Transformer Based Full-band and Sub-band Fusion Network for Speech Enhancement," 27 04 2021. [Online]. Available: https://arxiv.org/abs/2104.13002. [Accessed 12 01 2025].

[12] R. a. A. S. a. Y. B. Cao, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in Interspeech 2022, ISCA, 2022.

[13] J. J. Bosch, F. Fuhrmann and P. Herrera, "IRMAS: a dataset for instrument recognition in musical audio signals (1.0) [Data set]," in 13th International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal, 2014.

[14] C. J. J. Salamon, "A dataset and taxonomy for urban sound research," in 22nd ACM International Conference on Multimedia (ACM-MM'14), Orlando, FL, USA, 2014.

[15] Spotify, "pedalboard," github, [Online]. Available: https://github.com/spotify/pedalboard. [Accessed 04 02 2025].

[16] O. Ronnenberg, P. Fischer and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention (MICCAI), Munich, 2015.

[17] University of Freiburg, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Faculty of Engineering, 2015. [Online]. Available: https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/. [Accessed 04 02 2025].

[18] Y. Masuvama, K. Yatabe, Y. Koizumi, Y. Oikawa and N. Harada, "Deep Griffin–Lim Iteration," in IEEE International Conference on Acoustics, Speech, and Signal Processing, Brighton, 2019.

[19] Internation Telecommunications Union, "ITU-T P.862," 2001.

[20] International Telecommunications Union, "ITU-T P.800 : Methods for subjective determination of transmission quality," 1996.

[21] D. Stoller, S. Ewert and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," arXiv, 2018.

[22] I. Goodfellow, J. Pouget-Abadie, M. Mirza and B. Xu, "Generative Adversarial Networks," Advances in Neural Information Processing Systems , vol. 3, no. 11, 2014.

[23] K. Akesbi, "Audio Denoising for Robust Audio Fingerprinting (Master Thesis)," Paris-Saclay University, Orsay, France, 2022.

[24] M. A. Mohsin, "Deep Learning-based Noise Filtering for Speech Enhancement of Audio Signals," 01 2024. [Online]. Available: https://www.researchgate.net/publication/377388257_Deep_Learning-based_Noise_Filtering_for_Speech_Enhancement_of_Audio_Signals. [Accessed 12 01 2025].

[25] K. A. K. G. J. T. S. a. B. Y. S. Abdulatif, "AeGAN: Time-Frequency Speech Denoising via Generative Adversarial Networks,," in 28th European Signal Processing Conference (EUSIPCO), Amsterdam, Netherlands, 2020.

[26] jaxony, "U-Net Implementation in Pytorch," github, 2017. [Online]. Available: https://github.com/jaxony/unet-pytorch/tree/master. [Accessed 12 01 2025].

[27] AALTO, "5.9. The Griffin-Lim algorithm: Signal estimation from modified short-time Fourier transform," [Online]. Available: https://speechprocessingbook.aalto.fi/Modelling/griffinlim.html. [Accessed 04 02 2025].