*fractal and fractional*

MDPI

*Article*

# Data Augmentation for Deep Learning-Based Speech Reconstruction Using FOC-Based Methods

**Bilgi Görkem Yazgaç * and Mürvet Kırcı**

Department of Electronics and Communication Engineering, Istanbul Technical University, Istanbul 34469, Turkey; ucerm@itu.edu.tr

* Correspondence: yazgacb@itu.edu.tr

**Abstract:** Neural audio reconstruction is an important subtopic of Neural Audio Synthesis (NAS), which is a current emerging topic of modern Artificial Intelligence (AI) applications. The objective of a neural audio reconstruction model is to achieve a viable audio waveform from an audio feature representation that excludes the phase information. Since the data-dependent nature of such systems demands an increased quantity of data, methods of increasing the quantity of data for neural network training arise as a topic of substantial interest. Although the applications of data augmentation methods for classification tasks are well documented, there is still room for development for applications of such methods on signal synthesis tasks. Additionally, the Fractional-Order Calculus (FOC) framework provides possibilities for quality applications for the signal processing domain. Still, it is important to show that the methods based on the FOC framework can be applied to different application domains to show the capabilities of this framework. In this paper, FOC-based methods are applied to a speech dataset for data augmentation purposes to increase the audio reconstruction performance of a neural network, a spectral consistency-based neural audio reconstruction model called Deep Griffin-Lim Iteration (DeGLI), with respect to objective measures PESQ and STOI. An FOC-based method for rescaling linear frequency for augmenting magnitude spectrogram data is proposed. Furthermore, together with an FOC-based phase estimation method, it is shown that an augmentation strategy that has the objective of increased spectral consistency should be considered in data augmentation for audio reconstruction tasks. The test results reveal that this type of strategy increases the performance of a spectral consistency-based neural audio reconstruction model by over 13% for smaller depths.

**Keywords:** fractional-order calculus; neural audio reconstruction; data augmentation; spectral consistency

Academic Editor: António Lopes

Received: 26 December 2024
Revised: 13 January 2025
Accepted: 17 January 2025
Published: 21 January 2025

**Citation:** Yazgaç, B.G.; Kırcı, M. Data Augmentation for Deep Learning-Based Speech Reconstruction Using FOC-Based Methods. *Fractal Fract.* **2025**, *9*, 56. https://doi.org/10.3390/fractalfract9020056

## 1. Introduction

Fractional-Order Calculus (FOC) extends the concepts of differentiation and integration to non-integer orders, which traces back to discussions between Leibniz and L'Hospital in the 17th century [1]. By incorporating non-integer derivation orders, fractional calculus introduces an additional degree of flexibility, making it particularly effective in areas such as object modeling, performance optimization, and describing natural dynamic systems with memory [2]. Additionally, signal processing tools that are based on the FOC are also developed. FOC has a significant connection to fractal theory, which is also used for signal-processing applications [3]. For instance, assuming a stochastic signal following a well-defined fractal model, FOC-based methods can estimate the frequency characteristics of a signal [4]. Moreover, the model parameters derived within the fractal framework have applications in tasks such as signal texture segmentation [5]. Fractal theory can be used to describe local properties of signals, offering simplified geometrical or statistical descriptions regardless of whether the signal exhibits fractal properties [6].

In the literature, FOC-based models are used for reducing the number of linear prediction parameters. Differentiating a signal by an appropriate fractional order allows manipulation of its autocorrelation function, thereby reducing the number of linear prediction parameters needed. Improved signal prediction performance is documented in applications such as speech signal prediction, where fractional linear prediction methods based on weighted sums of fractional derivatives outperformed conventional techniques [7]. Given its non-local nature, fractional calculus is particularly suitable for handling signals with memory. The approaches that incorporate limited memory demonstrated both high prediction accuracy and a reduction in the number of linear prediction coefficients [8,9] required for audio signal encoding [10]. Similarly, the excitation in an autoregressive model of speech can be modeled with respect to the fractional derivatives of Gaussian noise [11]. Fractional derivatives can also serve as a framework for fractal analysis in audio processing with applications in speech recognition, voiced–unvoiced speech separation [12], and speaker emotion classification [13]. Notably, the fractal geometry-based features perform comparably to the Mel Frequency Cepstral Coefficients in speech classification tasks [14].

In image processing, FOC-based approaches are applied primarily through fractional differential masks, which are integral to edge detection algorithms [15]. The flexibility of tuning fractional derivative orders can enhance the performance of edge detection and segmentation filters. This methodology is successfully employed in diverse applications, including satellite image segmentation [16] and biomedical imaging, such as brain tomography analysis [17].

Recent works increased the applications of FOC-based methods on neural network-based signal processing approaches [18]. Especially in computer vision tasks, FOC approaches are used for denoising [19,20], medical image enhancement [21], and satellite and medical image segmentation [22,23]. Additionally, FOC-based approaches can be useful in accelerating the optimization of neural network training [24].

The speech synthesis problem is one of the most interested areas of research due to its effects on various engineering fields such as building interactive engineering products, audiobooks, navigation services, home automation products, or providing quality communication tools [25]. The purpose of speech synthesis is to produce a natural and intelligible waveform from a set of conditional variables.

The most significant application of speech synthesis procedures has been as a part of Text-to-Speech (TTS) applications. TTS applications provide frameworks for generating speech waveforms from text inputs [26]. These frameworks consist of two separate stages. Conventionally, the first stage is tasked with producing intermediate acoustic features from the input text. This stage is called the acoustic model. The second stage, which is also called a vocoder, then produces audio waveforms from the given representations [27]. Especially after the surge of deep learning, the models that are motivated to combine the two separate parts into one gained traction [28]. These types of approaches are called Neural Audio Synthesis (NAS) approaches. In the literature, the WaveNet architecture is often regarded as the benchmark generative model in the neural audio synthesis field [29]. WaveNet takes raw audio samples as input and models the joint probability of a waveform as a product of conditional probabilities [30]. In practical applications, WaveNet generates signals by conditioning on acoustic features, such as Mel Spectrograms [31]. Utilizing a Convolutional Neural Network (CNN) architecture, WaveNet defines causal convolutions to maintain causal output and employs dilated convolutions to reduce computational costs. This parallel processing approach, enabled by the CNN architecture, results in a faster system compared to Recurrent Neural Network (RNN) approaches [32]. The dilated convolution method also expands the receptive field of the network. By employing different optimization techniques, the WaveRNN model became one of the first sequential neural models capable of real-time audio synthesis on limited resource setups such as mobile phone CPUs [33]. Achieving competitive results with WaveNet, WaveRNN incorporated recurrent elements for a data-driven approach to audio synthesis. Essentially, WaveRNN consists of two neural networks: one models the most significant half of a 16-bit sampled speech signal, while the other models the least significant half [33]. This approach practically separates the tasks of estimating a signal's spectral shape and its stochastic elements. Generative Adversarial Networks (GANs) show promise in representing data features, making them applicable to audio synthesis problems. WaveGAN which processes waveforms as input and SpecGAN which uses spectrograms became the premier GAN-based models for audio synthesis [34].

A dominant number of audio-related applications dictate the analysis and modification of the Short-Time Fourier Transform (STFT) and the Short-Time Fourier Transform Magnitude (STFTM) representations of audio signals [35]. Audio enhancement [36], time and pitch modification [37], or reverberation analyses [38] are some examples of this procedure. In terms of TTS, transforming acoustic features to time-frequency representations, such as STFTM, is relatively easier than producing the waveform itself [28]. In such cases, the complex characteristics the phases of signals, are generally lost.

In the literature, it has been shown that an appropriate estimation of phase from the STFTM is possible [39]. This family of approaches is called audio reconstruction. Conventional methods for audio reconstruction are usually a member of the phase vocoder family. The basic phase vocoder method represents a signal as a combination of sine waves, where the key factors that need to be identified through analysis are the changing amplitude and frequency of each individual sine wave over time, which are present in the STFTM representation [40]. Additionally, spectral consistency-based approaches are used for audio reconstruction. The limited length of signal segments and the form of the spectral analysis window cause dependencies between the spectral coefficients of neighboring frequency bands, known as spectral redundancy which affects both spectral amplitudes and phases [36]. Exploiting the spectral redundancy, Griffin and Lim's Algorithm (GLA) estimates the spectral phases based on the spectral amplitudes of a speech signal with iterations [41]. In this method, the STFT and its inverse (ISTFT) are computed repeatedly while keeping the spectral amplitude fixed and only updating the phase. The STFTM-based phase reconstruction and iterative methods can be used together to improve audio reconstruction performance [42]. Because GLA and its derivatives are iterative algorithms, they are time-consuming [41,43]. Therefore, in areas where application speed is a concern, different algorithms have been proposed, such as Single-Pass Spectrogram Inversion (SPSI) [44]. The SPSI not only outputs applicable results but also provides a better initial phase estimate for iterative methods, such as GLA [44]. Recently, various methods have been proposed for non-iterative signal reconstruction problems that claimed improved results concerning SPSI [39,45].

The success of neural network approaches in audio synthesis is well-documented in the literature [46,47]. On the other hand, the focus of this research is intentionally limited to neural network models with a reduced number of parameters, in line with the hardware constraints of this study. In this study, a relatively humble PC setup that has NVIDA GeForce 1650Ti GPU is used. Furthermore, a network model with an especially smaller number of parameters is used for the experiments [48]. Since the smaller number of parameters of a network model indicates a limited performance, additional data augmentation methods have been employed to increase the capability of the neural network model.

This work aims to contribute to the literature as follows. Firstly, an FOC-based data augmentation method that works on STFT representations of data is introduced. Secondly, two data augmentation strategies are proposed. The first strategy employs only the proposed method which is similar to conventional data augmentation approaches for classification problems. The second strategy uses an additional FOC-based phase estimation procedure, aiming to create consistent spectrograms. It must be noted that the proposed strategies are based on FOC for both STFT representation augmentation and phase estimation, expanding the area of usage of fractional calculus. The experiments show the advantages of the second strategy for neural audio reconstruction problems. Thirdly, it is shown that a data augmentation strategy that has an objective of creating consistent spectrograms increases the evaluation performance of a neural audio reconstruction model with respect to the baseline implementation, providing an opportunity for smaller-sized network implementation.

This paper is organized as follows. In Section 2, in addition to discussing the concept of data augmentation, the proposed FOC-based data augmentation method that works on the linear spectrogram of a signal is introduced. Background information about the network that is used in this work, data augmentation strategies with or without a phase vocoding structure, and the dataset are also provided in this section as subsections. Section 3 contains information about the neural network implementation and neural audio reconstruction result comparisons of two different data augmentation strategies. An additional comparison of the data augmentation strategies in terms of spectral consistency is also provided in this section. Section 4 discusses the experiment results and provides conclusions about a data augmentation strategy that aims to produce consistent spectrograms. Section 5 summarizes the overall contributions of this work.

## 2. Materials and Methods

Since the data-dependent nature of deep learning systems demands an increased quantity of data, methods to increase the data quantity to train a neural network create a substantial interest [49]. Although the applications of data augmentation methods for classification tasks are well-documented [50], the number of applications of such methods on signal synthesis tasks is relatively lower.

Data augmentation is designed to expand the feature space while retaining the original labels of given data, thereby enhancing model performance and reducing overfitting [51,52]. For example, speech recognition systems frequently utilize artificially generated data [53,54]. Common applications of audio data augmentation in the time domain or time-frequency domain include noise addition, time stretching, time shifting, and pitch shifting [54]. Other methods involve warping the linear frequency scale during spectrogram creation to generate new data [55]. In many deep-learning audio applications, the log-Mel Spectrogram, which is a transformation of audio samples, is treated as an image-like input for neural networks. Consequently, data augmentation techniques developed for images, such as sparse image warping and masking, are adapted for audio applications, as seen in the SpecAugment strategy [55]. In [55], it is also shown that in addition to an augmentation method, the application policy or strategy is also important for successful results.

The popularity of computer vision-based deep learning has led to the development of diverse data augmentation strategies for images, such as flipping, rotation, cropping, color jittering, and edge enhancement. For example, Sobel operator-based edge enhancement has been effectively used in CNN-based image classification tasks [52]. Since the 2D spectrogram representation of an audio sample resembles an image, these augmentation methods can be directly applied to audio-related problems.

One of the widely used data augmentation methods in speech recognition is called Vocal Tract Length Perturbation (VTLP) [56]. By addressing the speaker variability caused by differences in vocal tract length, the linear frequency axis of an audio spectrogram is warped using a randomly selected warp factor derived from audio sample statistics. VTLP procedure produces a new spectrogram by applying different weights that are based on the warped frequency scale of the old spectrogram representation of the audio sample. Beyond speech recognition, VTLP is also applied to tasks such as animal audio classification [57] and environmental sound classification tasks [57]. Similar warping-based techniques have shown promising results in acoustic event detection problems [58].

### *2.1. Fractional Order Scaling*

Warping methods such as VTLP show that warping the frequency scale and increasing data size enhance a deep learning model's classification accuracy [56]. In this work, a method based on fractional-order differentiation is proposed for data augmentation purposes.

For fractional differ-integration of linear frequency scale, the Riemann–Liouville (RL) definition of fractional derivative is used [1]. It must be noted that a similar approach could be applied with using other definitions of fractional derivatives such as the Grünwald–Letnikov (GL) derivative. The numerical algorithm for the RL definition of the fractional derivative at point j [59,60] can be given as in Equation (1).

$$\left[\mathrm{D}^{\alpha}\mathrm{f}(\mathrm{x}_{\mathrm{j}})\right]_{\mathrm{RL}} = \mathrm{h}^{-\alpha}\sum_{\mathrm{k}=0}^{\mathrm{j}} A_{\mathrm{k,j}}\mathrm{f}(\mathrm{x}_{\mathrm{k}}) \tag{1}$$

The $A_{\mathrm{k,j}}$ parameters can be calculated as shown in Equation (2).

$$A_{\mathrm{k,j}} = \frac{1}{\Gamma(2-\alpha)}\begin{cases} (\mathrm{j}-1)^{1-\alpha} - (\mathrm{j}+\alpha-1)\mathrm{k}^{-\alpha}, & \mathrm{k}=0 \\ (\mathrm{j}-\mathrm{k}+1)^{1-\alpha} + (\mathrm{j}-\mathrm{k}-1)^{1-\alpha} - 2(\mathrm{j}-\mathrm{k})^{1-\alpha}, & 1 \le \mathrm{k} \le \mathrm{j}-1 \\ 1, & \mathrm{k}=1 \end{cases} \tag{2}$$

A lower triangular matrix can be produced by calculated $A_{\mathrm{k,j}}$ parameters. This matrix **R** can be seen in Equation (3).

$$\mathbf{R} = \frac{1}{\Gamma(2-\alpha)}\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ A_{0,1} & 1 & 0 & \cdots & 0 \\ A_{0,2} & A_{1,2} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{0,\mathrm{N}} & A_{1,\mathrm{N}} & A_{2,\mathrm{N}} & \cdots & 1 \end{bmatrix} \tag{3}$$

This numerical approach can be represented in a matrix multiplication form as in Equation (4), where **R** is a matrix that consists of $A_{\mathrm{k,j}}$ parameters as shown in Equation (3) and **f** is a vector that contains N + 1 function value of $\mathrm{f}(\mathrm{x})$.

$$[\mathrm{D}^{\alpha}\mathrm{f}(\mathrm{x})]_{\mathrm{RL}} = \mathrm{h}^{-\alpha}\mathbf{R}\cdot\mathbf{f} \tag{4}$$

In practice, the **f** vector contains frequency values that correspond to each frequency bin for a selected window size. Since every step represents a frequency bin number, the h value in (4) and the function step size can be taken as 1.

By taking the information above into consideration, a method named the Fractional-Order Frequency Scale is given in Equation (5). This approach enables the production of a corresponding value for each value on a frequency scale.

$$[\mathrm{D}^{\alpha}\mathrm{f}(\mathrm{x})]_{\mathrm{RL}} = \mathbf{R}\cdot\mathbf{f} \tag{5}$$

For $\alpha = -0.1$ and $\alpha = 0.1$ the process can be visualized as in Figure 1.

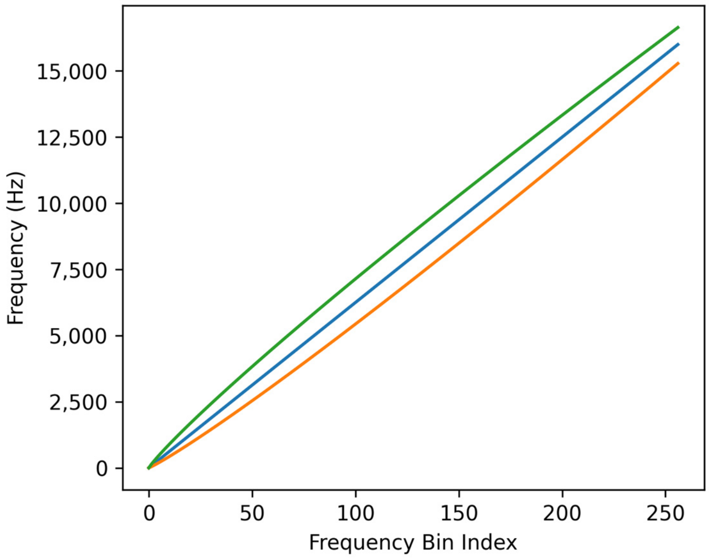


**Figure 1.** Rescaling of linear frequency scale by deriving with fractional order (Blue: $\alpha = -0.1$, Orange: $\alpha = 0$ (linear scale), Green: $\alpha = 0.1$).

Following the production of the new frequency scale, a new spectrogram is generated by applying different weights that are based on the new frequency scale of a given audio spectrogram. This mapping approach uses similar designs of mapping spectrograms on different psychoacoustic scales [56]. In practice, by applying a set of equally distanced and overlapping triangular filters on the new frequency scale, a set of weights is calculated to be applied to the complex coefficients of the STFT.

### *2.2. Data Augmentation Strategies*

Using the proposed method, two augmentation strategies depicted in Figure 2 are applied. Augmentation Strategy 1 is quite straightforward. Firstly, for each complex STFT matrix of audio samples, a new frequency scale is calculated with respect to a randomly given fractional order. Then, the audio sample is mapped and warped on a fractionally differentiated frequency scale to produce a complex augmented STFT matrix. Augmentation Strategy 2 differs from the first strategy with the added method to its output. The complex STFT matrix is mapped and warped on the fractionally derived frequency scale to produce a complex augmented STFT matrix as in the first strategy. Additionally, the amplitude of the complex augmented STFT matrix is calculated and the new phase structure is produced by estimating with the Fractional Differential Equation (FDE)-based phase estimation method [61] to produce a new augmented STFT. Proposed by the authors of this work, this method is shown to be capable of producing consistent spectrograms. Further information is provided in depth in [61]. Additionally, in a related work by the authors [62], the mapping procedure of spectrogram coefficients on a new frequency-like scale is also explained.

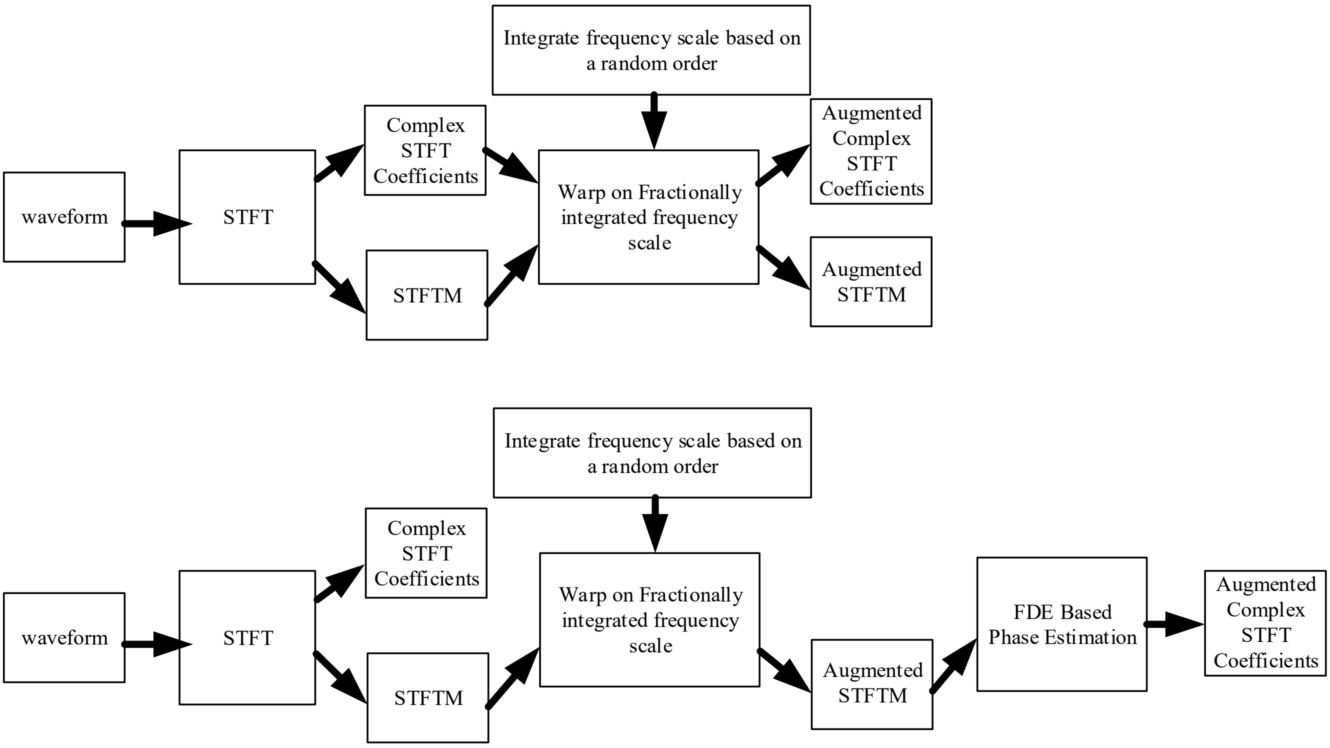

Integrate frequency scale based on a random order
waveform
STFT
Complex STFT Coefficients
STFTM
Warp on Fractionally integrated frequency scale
Augmented Complex STFT Coefficients
Augmented STFTM
Integrate frequency scale based on a random order
waveform
STFT
Complex STFT Coefficients
STFTM
Warp on Fractionally integrated frequency scale
Augmented STFTM
FDE Based Phase Estimation
Augmented Complex STFT Coefficients

**Figure 2.** Augmentation strategies: (**above**) Strategy 1, (**below**) Strategy 2.

### *2.3. Deep Griffin-Lim Iteration*

In this work, a recent approach for audio reconstruction with deep neural networks called Deep Griffin-Lim Iteration (DeGLI) is applied [48]. DeGLI is intentionally chosen because, the proposed implementation of this approach requires a smaller number of trainable parameters, resulting in less training time and the architecture enables a flexible implementation for evaluation in terms of network depth. In Table 1, trainable parameter comparisons of some NAS architectures are provided.

**Table 1.** Number of trainable parameters for some NAS architectures.

| Neural Audio Synthesis Architecture | Number of Trainable Parameters |
|---|---|
| Wavenet-30 [63] | 4.57 M |
| WaveRNN-896 [33] | 3 M |
| LPCNet [64] | 843 K–1.24 M |
| GlotNet [32] | 602 K–1.56 M |

DeGLI is a method for reconstructing audio signals that combines the GLA approach (Appendix A) with a deep neural network (DNN). To reconstruct the phase for a given STFTM representation, this approach uses a number of concatenated sub-blocks as shown in Figure 3.

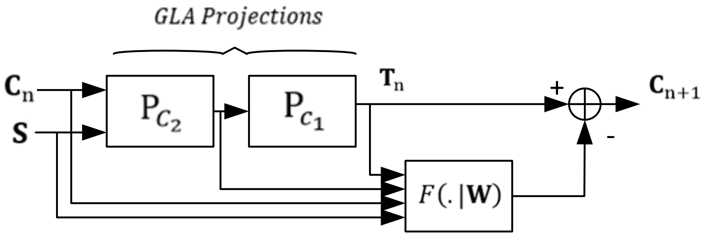


**Figure 3.** Block diagram view of a DeGLI sub-block:F( .|**W**) is a deep neural network with trainable parameters **W**, **S** is the amplitude spectrogram of a complex spectrogram **C**. $P_{c_1}$ (A8) and $P_{c_2}$ (A10) are matrix projections based on GLA.

The number of sub-blocks can be adjusted to control the computational cost at inference time. Owing to this design property, the DeGLI allows the same trained DNN to be used in different applications with varying computational requirements. Similar approaches for defining iterative algorithms as neural networks are called deep unfolding [65]. This approach enables the user to decide on sub-block size for inference.

Figure 4 shows the implementation of DeGLI for N iterations.

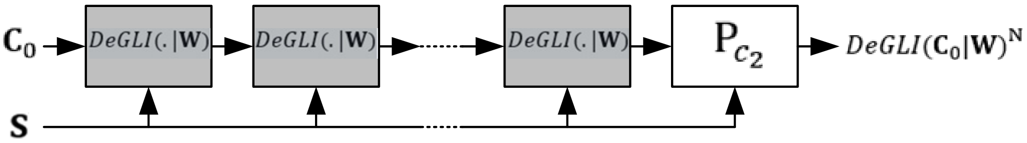


**Figure 4.** DeGLI implementation for N iterations.

A unique feature of DeGLI is the training strategy for the DNN, which is specifically designed for iterative use. The DNN is trained on a denoising task with a single sub-block, which reduces the memory required for training and improves stability compared to end-to-end training. A similar approach in the literature is called Plug-and-Play (PnP). In PnP, arbitrary denoiser models are applied to increase the training efficiency [66]. More in-depth information about the DeGLI architecture can be found in [48].

### *2.4. Dataset*

In this work, the Texas Instruments/Massachusetts Institute of Technology (TIMIT) dataset is used. TIMIT is a read speech corpus, which has been used for benchmarking speech processing implementations [67]. The corpus contains 16-bit, 16 kHz speech samples from various dialects of American English from male and female participants. This dataset contains training and test subsets. Since this dataset provides gender and dialect variabilities it is preferred for this work. In comparison to the datasets such as LJSpeech, TIMIT is sampled with a smaller sampling frequency [48]. This property enables the implementation of DeGLI in a reduced size without the need for downsampling or upsampling the training data.

## 3. Results

Due to hardware constraints of the experiment setup, actions are taken to reduce network parameter size. For this purpose, the channel size of Complex Convolutional Layers [48] is reduced from 64 to 32, resulting in nearly four times smaller trainable network parameters. Additionally, two training parameters, batch size, and the number of epochs for training are also reduced. The reduced batch size enables using less memory and the reduced epoch number results in shorter training time. Lastly, the learning rate is kept the same for the duration of training. The implemented network for the present experiments and its training parameters are selected as given in Table 2, in comparison to the original DeGLI implementation.

**Table 2.** Implemented network architecture and training parameters in comparison with original DeGLI implementation.

| Architecture and Training Parameters | Original DeGLI [48] | Implemented DeGLI |
|---|---|---|
| # of Amplitude Informed Gated Convolutional (AI-GC) Layers | 3 | 3 |
| # of Complex Convolutional Layers | 1 | 1 |
| # of Channels | 64 | 32 |
| Filter size of AI-GC | 5 × 3 | 5 × 3 |
| Filter size of Last Complex Convolutional Layer | 1 × 1 | 1 × 1 |
| Stride for Convolutional Layers | 1 × 1 | 1 × 1 |
| # of Trainable Parameters | 380 k | 98 k |
| Optimizer | ADAM | ADAM |
| Initial Learning Rate Step Size | 0.0004 | 0.0004 |
| Batch Size | 32 | 16 |
| # of Epochs | 300 | 100 |
| Randomly selected SNR values for Denoiser Training | $[-6, 12]$ dB | $[-6, 12]$ dB |

Additionally, similar reasons that lead to implementing a smaller network also dictate the use of a different dataset than the one used in the original DeGLI paper. The original DeGLI paper employs the LJSpeech dataset [48]; however, the TIMIT dataset is used for training in this work. The implemented DeGLI block is trained on 16 kHz sampled TIMIT samples, each with 1 s of duration. The window length for STFT and ISTFT is 512 and the hop length is 128.

In this work, audio reconstruction results of two different augmentation strategies are compared with respect to an objective measure called Perceptual Evaluation of Speech Quality (PESQ) [68]. The PESQ is a correlated measure with human audio perception. A dataset of 11,071 audio samples is augmented with respect to an arbitrarily chosen fractional derivative order from a set of $[-0.1, -0.05, 0.05, 0.1]$. Each audio sample is augmented with respect to randomly selected two-order values from the given set. As a result, when combined with the original data, a three times larger dataset is produced for testing with each strategy. A DeGLI sub-block is trained as a denoiser. The Gaussian Noise is added to each training data sample to produce noisy inputs between $[-6, 12]$ dB. Using an L2 loss [48], the neural network is tasked to denoise the noisy input samples.

In the first test, the effect of Augmentation Strategy 1 is compared with the audio reconstruction using a baseline DeGLI model that is trained on a non-augmented dataset. The result can be seen in Figure 5. It shows that the direct application of fractional-order scaling on complex STFT, a strategy that can be applied to classification problems, reduces the capability of the DeGLI network.

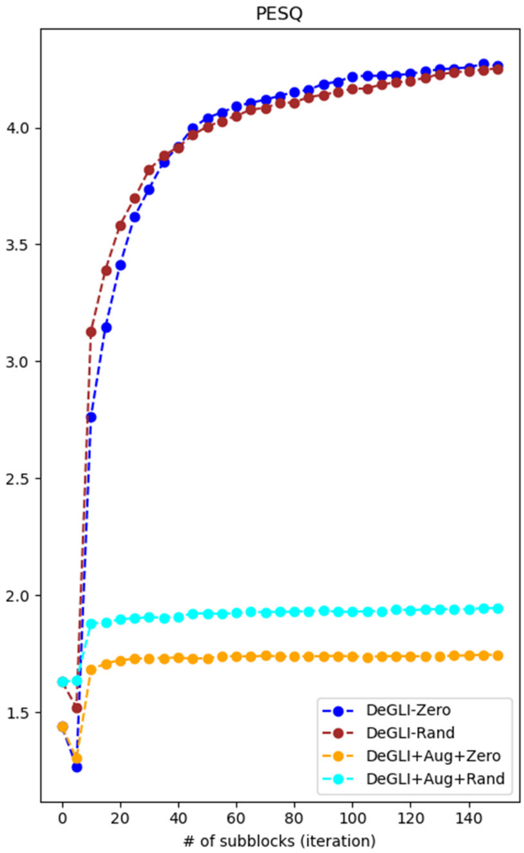


**Figure 5.** Test of Augmentation Strategy 1 with orders [−0.1, −0.05, 0.05, 0.1] on TIMIT, in terms of neural network depth.

In Figure 5, the "DeGLI + Zero" curve shows the PESQ result when the complex component of the spectrogram is initialized as 0 and the non-augmented dataset is used for neural network model training. The "DeGLI + Rand" curve represents the results when the complex component of the spectrogram is initialized randomly and the non-augmented dataset is used for model training. "DeGLI + Aug + Zero" and "DeGLI + Aug + Rand" are results for 0 and random initializations on a network trained on the augmented dataset.

This result outlines the inapplicability of such an augmentation strategy on a neural network model that takes spectral consistency into account. To further analyze the reasons for this result, Figure 6 shows the Log-Spectral Convergence (A6) measures for two sets of augmented TIMIT datasets.

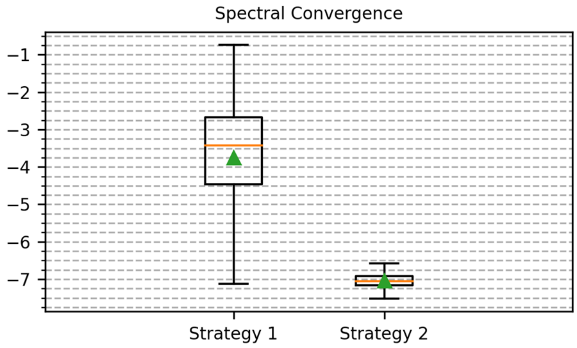
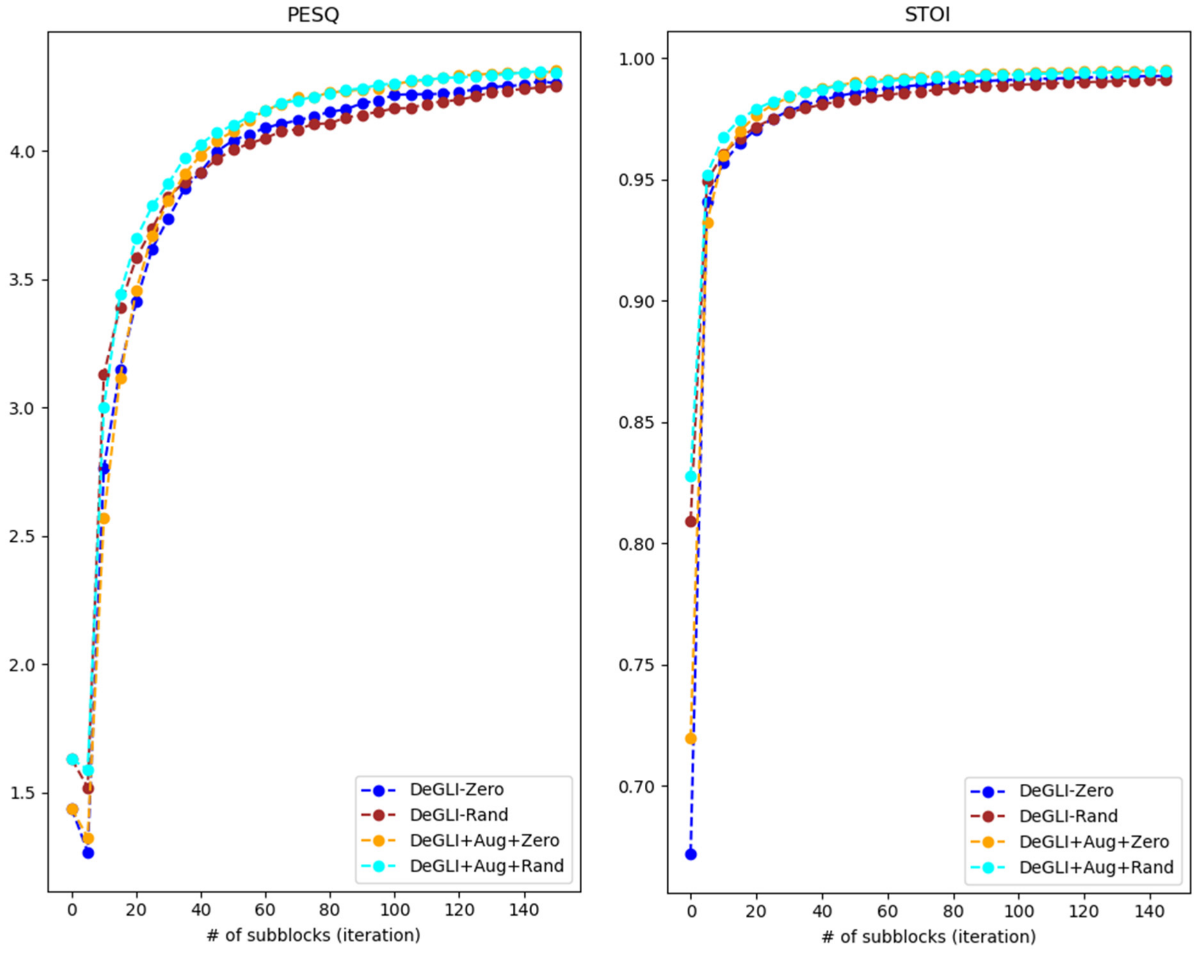


**Figure 6.** Log-Spectral Convergence comparison of Augmentation Strategy 1 and Augmentation Strategy 2.

The measure in Figure 6 is calculated by one iteration result of GLA on 30% of the training dataset. This result shows that using Strategy 2 with the added vocoder structure creates a more consistent spectrogram after augmenting a sample.

Repeating the audio reconstruction test with Augmentation Strategy 2 resulted in an increased reconstruction performance as seen in Figure 7a,b. The curve names have the same meanings as in Figure 5.



(**a**) (**b**)

**Figure 7.** Test of Augmentation Strategy 2 with orders [−0.1, −0.05, 0.05, 0.1] on TIMIT, in terms of neural network depth. (**a**) PESQ result (**b**) STOI result. Higher PESQ and STOI values indicate better performance.

From the results, it can be seen that augmenting the dataset three times with Augmentation Strategy 2 substantially increases the reconstruction performance of DeGLI not only in terms of PESQ but also STOI. The novel method proposed in this work successfully increases the performance of the neural network-based audio reconstruction model for both zero and random phase initialization.

The comparison of loss function values of DeGLI sub-block training for each training configuration also indicates the success of the Augmentation Strategy 2 as seen in Figure 8.

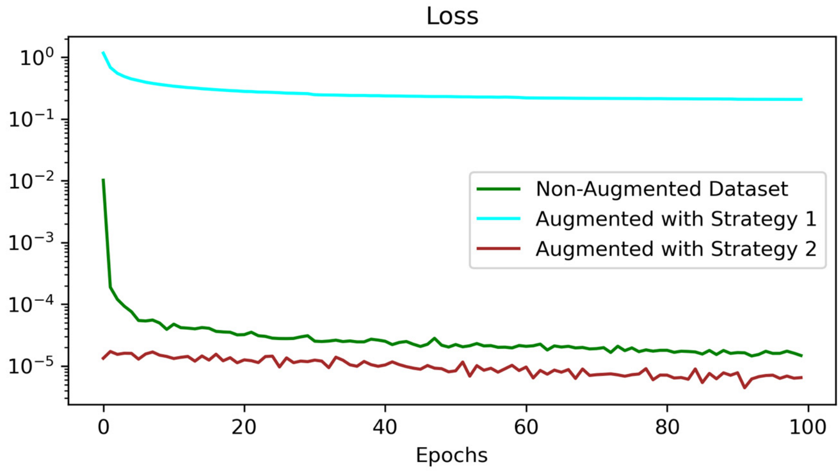


**Figure 8.** Training losses of DeGLI subblock in a denoiser mode for non-augmented dataset, augmented dataset with Strategy 1, and augmented dataset with Strategy 2.

In Figure 8, it can be seen that the training of a sub-block as a denoiser with Augmentation Strategy 1 increases the L2-loss of training. Since the number of batches is higher than the non-augmented dataset test, the loss curves for both augmentation tests are less steep than the test with the original dataset. As a result, Augmentation Strategy 2 produced both a less steep and reduced loss curve.

## 4. Discussion

In this work, two data augmentation strategies that employ the proposed Fractional Order Frequency Scale method are tested in terms of the audio reconstruction quality. For the tests, the dataset size is increased three times by augmenting each sample with different values of derivation order $\alpha$. The important differences between these two strategies are as follows: Strategy 1 directly uses the complex spectrogram and augments it by applying the data augmentation method; Strategy 2 calculates the amplitude spectrogram of augmented data, and by employing the FOC-based phase estimation method reproduces new data. The experiments with Strategy 1 produced catastrophic results in terms of reconstruction quality. This is due to its effect on spectral consistency. Comparing two augmentation methods in terms of log-SC shows that Strategy 2 is capable of producing more spectrally consistent data. The difference in spectral convergence results can be expected. In Augmentation Strategy 1, the resulting augmented complex STFT representation has a different STFTM but the same phase information, with respect to the original data. This causes an increased spectral redundancy for the augmented data. Since the classification problems mostly exclude phase (complex) information from data, Augmentation Strategy 1 can be useful in such tasks. On the contrary, the phase information is calculated from the augmented STFTM representation for Augmentation Strategy 2 using an FOC-based phase estimation method. In the training stage of the neural network, using the augmented dataset with Strategy 2, the denoiser model learns to produce consistent spectrograms. As a result, this novel strategy increases the audio spectrogram reconstruction quality of the implemented DeGLI model by up to 13.4% (10 subblocks) for a smaller number of sub-blocks, while using randomly assigned initial complex coefficients in terms of PESQ and STOI. The increase, in terms of STOI is relatively small. The experiment results give an idea about the inclusion of phase information in data augmentation approaches and provide an intuition for the capabilities of data augmentation methods that produce more spectrally consistent augmented data for training audio synthesis models. It must be noted that the DeGLI is designed to leverage spectral consistency. For future works, to further understand the general performance of the proposed data augmentation strategies, NAS architectures that use features that are derived from STFTM, e.g., Mel spectrograms, should be considered.

The capability of increased performance for smaller network sizes provides an advantage of fewer computation resources in the implementation of the audio reconstruction network. Additionally, a smaller-sized network makes it possible for implementations on limited hardware resources.

## 5. Conclusions

This work aims to provide three contributions to the literature. Firstly, an FOC-based data augmentation method that works on STFT representations of data is introduced. In relation to the similar approaches from the literature, this method can have possible applications for classification problems. On the other hand, this work focuses on the neural audio reconstruction task to understand the limitations of a warping-based data augmentation method for such problems. This objective provides the second contribution. To understand the applicability of the proposed method to a specific neural audio reconstruction problem, two data augmentation strategies are proposed. The first strategy employs only the fractional order scaling-based method which is similar to conventional data augmentation approaches for classification problems. In the second strategy, an additional FOC-based phase estimation procedure is used with the purpose of creating consistent spectrograms. In addition to expanding the application domains of FOC-based methods, the experiments show the advantage of a strategy that aims to create consistent spectrograms for neural audio reconstruction problems. Lastly, by increasing the evaluation performance of the neural network implementation in terms of widely accepted objective psychoacoustic measures, namely PESQ and STOI, especially for smaller depths, the proposed data augmentation strategy enables the implementation with a smaller-sized network using reduced computation resources.

**Author Contributions:** Conceptualization, B.G.Y.; methodology, B.G.Y.; software, B.G.Y.; validation, B.G.Y. and M.K.; resources, B.G.Y.; writing—original draft, B.G.Y.; writing—review & editing, B.G.Y. and M.K.; supervision, M.K.; project administration, M.K.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Scientific Research Projects Department of Istanbul Technical University, Project Number: MDK-2020-42479.

**Data Availability Statement:** The data used in this study is derived from the following resources available in the public domain: [https://www.kaggle.com/datasets/nltkdata/timitcorpus], accessed on 16 January 2025.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## Appendix A. Spectral Consistency

The STFT is a commonly used technique for analyzing signals in the time-frequency domain. However, it has some inherent limitations due to the windowing process involved, which can result in redundant information in the spectrogram. The spectral redundancy means that the coefficients obtained from the STFT may not necessarily form a valid spectrogram [41]. Spectral redundancy occurs because the windowing process in the STFT introduces overlaps between adjacent windows, causing redundant information in the frequency domain. This redundancy can lead to multiple possible signals having the same STFT coefficients, making it impossible to uniquely reconstruct the original signal from the STFT alone.

The problem of reconstructing a signal can be formulated as constructing a signal from a non-valid STFT magnitude. The problem can be expressed as finding a signal $\mathbf{x}^* \in \mathrm{R}^{\mathrm{L}}$ from a given set of non-negative coefficients **S**, such that the magnitude of the STFT of $\mathbf{x}^*$, $|\mathrm{G}\mathbf{x}|$, is as close as possible to **S** [69]. L represents the dimension of the space.

As a measure of closeness, the L2-norm provides a sufficient measure. The mathematical formulation for the problem in terms of optimization can be described as a minimization problem. Since G is a frame-dependent Gabor-based transform and $\mathbf{S} = |\mathbf{S}|$ is the real positive coefficients, the problem of finding a signal $\mathbf{x}^*$ that has a valid spectrogram can be defined in the following form (A1) [41].

$$\text{minimize}_{\mathbf{x}\in\mathrm{R}^{\mathrm{L}}} \| |\mathrm{G}\mathbf{x}| - \mathbf{S} \|_2 \tag{A1}$$

The problem can be translated as **S** is a valid STFT magnitude if there exists an **x** such that $|\mathrm{G}\mathbf{x}| = \mathbf{S}$. For consistency with optimization problem definitions, the problem can be defined with an optimization variable on the coefficient side. Here **C** corresponds to the complex coefficients of a spectrogram.

$$\text{minimize}_{\mathbf{C}\in\mathrm{C}^{\mathrm{M}\times\mathrm{N}}} \| |\mathbf{C}| - \mathbf{S} \|_2 \text{ s.t. } \exists \mathbf{x} \in \mathrm{R}^{\mathrm{L}} \Big| \mathbf{C} = \mathrm{G}\mathbf{x} \tag{A2}$$

The measure of error for the problem in (A2) is given in the form (A3).

$$\mathrm{E}(\mathbf{x}) = \frac{\| |\mathrm{G}\mathbf{x}| - \mathbf{S} \|_2}{\|\mathbf{S}\|_2} \tag{A3}$$

This error measure can be represented in the form of Spectral Signal to Noise Ratio (SSNR) as in (A4).

$$\mathrm{SSNR}(\mathbf{x}) = -10\log_{10}(\mathrm{E}(\mathbf{x})) \tag{A4}$$

Another representation of this measure is the Spectral Convergence (SC) as given in the Equation (A5) [43]. The SC is one of the most used objective speech quality metrics.

$$\mathrm{E}(\mathbf{x}) = \frac{\| \mathbf{S} - |\mathrm{G}\mathbf{x}| \|_2}{\|\mathbf{S}\|_2} \tag{A5}$$

The log-SC can be calculated as (A6).

$$\text{log-SC} = 10\log(\mathrm{E}(\mathbf{x})) \tag{A6}$$

The GLA is a double projection algorithm. It employs iterative projections of signal on set $\mathrm{C}_1$, which is the set of the admissible points of the optimization problem (A2) and set $\mathrm{C}_2$, which is the set of coefficients minimizing the optimization problem (A2). $\mathrm{C}_1$ and $\mathrm{C}_2$ constraints are sets that are in $\mathrm{C}^{\mathrm{M}\times\mathrm{N}}$ [41]. Here, M corresponds to the number of frequency channels, and N corresponds to the number of time indexes.

Since $\mathrm{C}_1$ is the set of admissible points for the problem as given in (A7), it is a hard constraint. It corresponds to the set of coefficients **C** that can be reached from the solution $\mathbf{x}^* \in \mathrm{R}^{\mathrm{L}}$ by applying transform G.

$$\mathrm{C}_1 = \left\{ \mathbf{C} \Big| \exists \mathbf{x} \in \mathrm{R}^{\mathrm{L}} \text{ s.t. } \mathbf{C} = \mathrm{G}\mathbf{x} \right\} \tag{A7}$$

The constraint $\mathrm{C}_1$ forces the solution to satisfy the consistency criterion. The projection can be defined as two transforms, ISTFT and STFT [42]. In Equation (A8), STFT is denoted as G and its pseudo inverse ISTFT is denoted as $\mathrm{G}^{\dagger}$.

$$\mathrm{P}_{\mathrm{c}_1}(\mathbf{C}) = \mathrm{G}\mathrm{G}^{\dagger}\mathbf{C} \tag{A8}$$

The $C_2$ constraint can be defined as in (A9).

$$C_2 = \left\{ \mathbf{C} \in C^{MxN} \middle| |\mathbf{C}| = \mathbf{S} \right\} \quad \text{(A9)}$$

This constraint forces non-negative coefficients **S** to be equivalent to the coefficients **C** that are in the set $C_1$. This soft constraint can be met with the following projection onto $C_2$ as in (A10).

$$P_{c_2}(\mathbf{C}) = \mathbf{S} \cdot e^{j\angle \mathbf{C}} \quad \text{(A10)}$$

The GLA can now be formulated as shown in the Algorithm A1.

**Algorithm A1** Griffin–Lim Algorithm [41]

```
Fix the initial phase $\angle \mathbf{C}_0$
Initialize $\mathbf{C}_0 = \mathbf{S} \cdot e^{\cdot j\angle \mathbf{C}}$
Iterate for n = 1, 2, . . . do
        $\mathbf{C}_n = P_{c_1}(P_{c_2}(\mathbf{C}_{n-1}))$
Until convergence
$\mathbf{x}^* = G^{\dagger}\mathbf{C}_n$
```

The iterative process of GLA aims to increase the spectral consistency to approximate a valid spectrogram for a given audio waveform.

## References

1. Podlubny, I. *Fractional Differential Equations: Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of Their Solution and Some of Their Applications*; Academic Press: Cambridge, MA, USA, 1999.
2. Petráš, I. *Fractional-Order Nonlinear Systems: Modeling, Analysis and Simulation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2011.
3. Ortigueira, M.; Machado, J. Which Derivative? *Fractal Fract.* **2017**, *1*, 3. [CrossRef]
4. Sabanal, S.; Nakagawa, M. The Fractal Properties of Vocal Sounds and Their Application in the Speech Recognition Model. *Chaos Solitons Fractals* **1996**, *7*, 1825–1843. [CrossRef]
5. Al-Akaidi, M. *Fractal Speech Processing*; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
6. Lévy-Véhel, J. Fractal Approaches in Signal Processing. *Fractals* **1995**, *3*, 755–775. [CrossRef]
7. Assaleh, K.; Ahmad, W.M. Modeling of Speech Signals Using Fractional Calculus. In Proceedings of the 2007 9th International Symposium on Signal Processing and Its Applications, Sharjah, United Arab Emirates, 12–15 February 2007; IEEE: Piscataway, NJ, USA, 2007; pp. 1–4. [CrossRef]
8. Despotovic, V.; Skovranek, T.; Peric, Z. One-Parameter Fractional Linear Prediction. *Comput. Electr. Eng.* **2018**, *69*, 158–170. [CrossRef]
9. Skovranek, T.; Despotovic, V.; Peric, Z. Optimal Fractional Linear Prediction with Restricted Memory. *IEEE Signal Process. Lett.* **2019**, *26*, 760–764. [CrossRef]
10. Skovranek, T.; Despotovic, V. Audio Signal Processing Using Fractional Linear Prediction. *Mathematics* **2019**, *7*, 580. [CrossRef]
11. Maragos, P.; Young, K.L. Fractal Excitation Signals for CELP Speech Coders. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Albuquerque, NM, USA, 3–6 April 1990; IEEE: Piscataway, NJ, USA, 1990; pp. 669–672. [CrossRef]
12. Maragos, P.; Potamianos, A. Fractal Dimensions of Speech Sounds: Computation and Application to Automatic Speech Recognition. *J. Acoust. Soc. Am.* **1999**, *105*, 1925–1932. [CrossRef] [PubMed]
13. Tamulevičius, G.; Karbauskaitė, R.; Dzemyda, G. Speech Emotion Classification Using Fractal Dimension-Based Features. *Nonlinear Anal. Model. Control* **2019**, *24*, 679–695. [CrossRef]
14. Pitsikalis, V.; Maragos, P. Analysis and Classification of Speech Signals by Generalized Fractal Dimension Features. *Speech Commun.* **2009**, *51*, 1206–1223. [CrossRef]
15. Mathieu, B.; Melchior, P.; Oustaloup, A.; Ceyral, C. Fractional Differentiation for Edge Detection. *Signal Process.* **2003**, *83*, 2421–2432. [CrossRef]
16. Henriques, M.; Valério, D.; Gordo, P.; Melicio, R. Fractional-Order Colour Image Processing. *Mathematics* **2021**, *9*, 457. [CrossRef]

17. Padlia, M.; Sharma, J. Brain Tumor Segmentation from MRI Using Fractional Sobel Mask and Watershed Transform. In Proceedings of the 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC), Indore, India, 17–19 August 2017; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6. [CrossRef]
18. Coelho, C.; Costa, M.F.P.; Ferrás, L.L. Fractional Calculus Meets Neural Networks for Computer Vision: A Survey. *AI* **2024**, *5*, 1391–1426. [CrossRef]
19. Bai, Y.-C.; Zhang, S.; Chen, M.; Pu, Y.-F.; Zhou, J.-L. A Fractional Total Variational CNN Approach for SAR Image Despeckling. In Proceedings of the Intelligent Computing Methodologies: 14th International Conference, ICIC 2018, Wuhan, China, 15–18 August 2018; pp. 431–442. [CrossRef]
20. Jia, X.; Liu, S.; Feng, X.; Zhang, L. Focnet: A Fractional Optimal Control Network for Image Denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6047–6056. [CrossRef]
21. Krouma, H.; Ferdi, Y.; Taleb-Ahmedx, A. Neural Adaptive Fractional Order Differential Based Algorithm for Medical Image Enhancement. In Proceedings of the 2018 International Conference on Signal, Image, Vision and their Applications (SIVA), Guelma, Algeria, 26–27 November 2018. [CrossRef]
22. Arora, S.; Suman, H.K.; Mathur, T.; Pandey, H.M.; Tiwari, K. Fractional Derivative Based Weighted Skip Connections for Satellite Image Road Segmentation. *Neural Netw.* **2023**, *161*, 142–153. [CrossRef] [PubMed]
23. Lakra, M.; Kumar, S. A Fractional-Order PDE-Based Contour Detection Model with CeNN Scheme for Medical Images. *J. Real-Time Image Process.* **2022**, *19*, 147–160. [CrossRef]
24. Chen, Y.; Wu, Z.; Lu, Y.; Chen, Y.; Wang, Y. Accelerated Gradient Descent Driven by Lévy Perturbations. *Fractal Fract.* **2024**, *8*, 170. [CrossRef]
25. Wagner, P.; Beskow, J.; Betz, S.; Edlund, J.; Gustafson, J.; Eje Henter, G.; Le Maguer, S.; Malisz, Z.; Székely, É.; Tånnander, C.; et al. Speech Synthesis Evaluation—State-of-the-Art Assessment and Suggestion for a Novel Research Program. In Proceedings of the 10th ISCA Workshop on Speech Synthesis (SSW 10), Vienna, Austria, 20–22 September 2019; ISCA: Singapore, 2019; pp. 105–110. [CrossRef]
26. Kaur, N.; Singh, P. Conventional and Contemporary Approaches Used in Text to Speech Synthesis: A Review. *Artif. Intell. Rev.* **2023**, *56*, 5837–5880. [CrossRef]
27. Shi, Z. A Survey on Audio Synthesis and Audio-Visual Multimodal Processing. *arXiv* **2021**, arXiv:2108.00443.
28. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards End-to-End Speech Synthesis. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; ISCA: Singapore, 2017; pp. 4006–4010. [CrossRef]
29. AlBadawy, E.A.; Gibiansky, A.; He, Q.; Wu, J.; Chang, M.-C.; Lyu, S. Vocbench: A Neural Vocoder Benchmark for Speech Synthesis. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; Volume 2022-May, pp. 881–885. [CrossRef]
30. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**, arXiv:1609.03499.
31. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 5–20 April 2018; IEEE: Piscataway, NJ, USA, 2018; Volume 2018-April, pp. 4779–4783. [CrossRef]
32. Juvela, L.; Bollepalli, B.; Tsiaras, V.; Alku, P. GlotNet—A Raw Waveform Model for the Glottal Excitation in Statistical Parametric Speech Synthesis. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1019–1030. [CrossRef]
33. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; van den Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient Neural Audio Synthesis. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
34. Donahue, C.; McAuley, J.; Puckette, M. Adversarial Audio Synthesis. In Proceedings of the International Conference on Learning Representations (ICLR) 2019, New Orleans, LA, USA, 6–9 May 2019; pp. 1–16.
35. Griffin, D.; Lim, J. Signal Estimation from Modified Short-Time Fourier Transform. In Proceedings of the ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing, Boston, MA, USA, 14–16 April 1983; IEEE: Piscataway, NJ, USA, 1983; Volume 8, pp. 804–807. [CrossRef]
36. Krawczyk, M.; Gerkmann, T. STFT Phase Reconstruction in Voiced Speech for an Improved Single-Channel Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1931–1940. [CrossRef]
37. Stefanakis, N.; Abel, M.; Bergner, A. Sound Synthesis Based on Ordinary Differential Equations. *Comput. Music J.* **2015**, *39*, 46–58. [CrossRef]
38. Laroche, J.; Dolson, M. Phase-Vocoder: About This Phasiness Business. In Proceedings of the 1997 Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 19–22 October 1997. [CrossRef]

39. Prusa, Z.; Balazs, P.; Sondergaard, P.L. A Noniterative Method for Reconstruction of Phase from STFT Magnitude. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1154–1164. [CrossRef]
40. Dolson, M. The Phase Vocoder: A Tutorial. *Comput. Music J.* **1986**, *10*, 14. [CrossRef]
41. Perraudin, N.; Balazs, P.; Sondergaard, P.L. A Fast Griffin-Lim Algorithm. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 1–4. [CrossRef]
42. Le Roux, J.; Kameoka, H.; Ono, N.; Sagayama, S. Fast Signal Reconstruction from Magnitude Stft Spectrogram Based on Spectrogram Consistency. In Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10), Graz, Austria, 6–10 September 2010; pp. 397–403.
43. Masuyama, Y.; Yatabe, K.; Oikawa, Y. Griffin–Lim Like Phase Recovery via Alternating Direction Method of Multipliers. *IEEE Signal Process. Lett.* **2019**, *26*, 184–188. [CrossRef]
44. Beauregard, G.T.; Harish, M.; Wyse, L. Single Pass Spectrogram Inversion. In Proceedings of the 2015 IEEE International Conference on Digital Signal Processing (DSP), Singapore, 21–24 July 2015; IEEE: Piscataway, NJ, USA, 2015; Volume 2015-September, pp. 427–431. [CrossRef]
45. Prusa, Z.; Søndergaard, P.L. Real-Time Spectrogram Inversion Using Phase Gradient Heap Integration. In Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16), Brno, Czech Republic, 5–9 September 2016; pp. 17–21.
46. Valin, J.M.; Skoglund, J. LPCNET: Improving Neural Speech Synthesis Through Linear Prediction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; Volume 2019-May, pp. 5891–5895. [CrossRef]
47. Govalkar, P.; Fischer, J.; Zalkow, F.; Dittmar, C. A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction. In Proceedings of the 10th ISCA Workshop on Speech Synthesis (SSW 10), Vienna, Austria, 20–22 September 2019; ISCA: Singapore, 2019; pp. 7–12. [CrossRef]
48. Masuyama, Y.; Yatabe, K.; Koizumi, Y.; Oikawa, Y.; Harada, N. Deep Griffin–Lim Iteration: Trainable Iterative Phase Reconstruction Using Neural Network. *IEEE J. Sel. Top. Signal Process.* **2021**, *15*, 37–50. [CrossRef]
49. Halevy, A.; Norvig, P.; Pereira, F. The Unreasonable Effectiveness of Data. *IEEE Intell. Syst.* **2009**, *24*, 8–12. [CrossRef]
50. Shorten, C.; Khoshgoftaar, T.M. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
51. Perez, L.; Wang, J. The Effectiveness of Data Augmentation in Image Classification Using Deep Learning. *arXiv* **2017**, arXiv:1712.04621.
52. Taylor, L.; Nitschke, G. Improving Deep Learning with Generic Data Augmentation. In Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 18–21 November 2018; pp. 1542–1547. [CrossRef]
53. Ragni, A.; Knill, K.M.; Rath, S.P.; Gales, M.J.F. Data Augmentation for Low Resource Languages. In *Interspeech 2014*; ISCA: Singapore, 2014; Volume 2019-September, pp. 810–814. [CrossRef]
54. Rebai, I.; Benayed, Y.; Mahdi, W.; Lorré, J.P. Improving Speech Recognition Using Data Augmentation and Acoustic Model Fusion. *Procedia Comput. Sci.* **2017**, *112*, 316–322. [CrossRef]
55. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; ISCA: Singapore, 2019; pp. 2613–2617. [CrossRef]
56. Jaitly, N.; Hinton, G.E. Vocal Tract Length Perturbation (VTLP) Improves Speech Recognition. In Proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language, Marseille, France, 22–23 August 2013; Volume 90, pp. 42–51.
57. Nanni, L.; Maguolo, G.; Paci, M. Data Augmentation Approaches for Improving Animal Audio Classification. *Ecol. Inform.* **2020**, *57*, 101084. [CrossRef]
58. Nam, H.; Kim, S.-H.; Park, Y.-H. Filteraugment: An Acoustic Environmental Data Augmentation Method. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 4308–4312. [CrossRef]
59. Adams, M. Differint: A Python Package for Numerical Fractional Calculus. *arXiv* **2019**, arXiv:1912.05303.
60. Diethelm, K. An Algorithm for the Numerical Solution of Differential Equations of Fractional Order. *Electron. Trans. Numer. Anal.* **1997**, *5*, 1–6.
61. Yazgaç, B.G.; Kırcı, M. Fractional Differential Equation-Based Instantaneous Frequency Estimation for Signal Reconstruction. *Fractal Fract.* **2021**, *5*, 83. [CrossRef]
62. Yazgaç, B.G.; Kırcı, M. Fractional-Order Calculus-Based Data Augmentation Methods for Environmental Sound Classification with Deep Learning. *Fractal Fract.* **2022**, *6*, 555. [CrossRef]
63. Ping, W.; Peng, K.; Zhao, K.; Song, Z. WaveFlow: A Compact Flow-Based Model for Raw Audio. *arXiv* **2019**, arXiv:1912.01219.
64. Popov, V.; Kudinov, M.; Sadekova, T. Gaussian Lpcnet for Multisample Speech Synthesis. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; IEEE: Piscataway, NJ, USA, 2020; Volume 2020-May, pp. 6204–6208. [CrossRef]

65. Hershey, J.R.; Roux, J.L.; Weninger, F. Deep Unfolding: Model-Based Inspiration of Novel Deep Architectures. *arXiv* **2014**, arXiv:1409.2574.
66. Venkatakrishnan, S.V.; Bouman, C.A.; Wohlberg, B. Plug-and-Play Priors for Model Based Reconstruction. In Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing, Austin, TX, USA, 3–5 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 945–948. [CrossRef]
67. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Pallett, D.S.; Dahlgren, N.L.; Zue, V.; Fiscus, J.G. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*; NIST Speech Disc 1-1.1; Linguistic Data Consortium: Philadelphia, PA, USA, 1993.
68. Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC Press: Boca Raton, FL, USA, 2017.
69. Le Roux, J.; Ono, N.; Sagayama, S. Explicit Consistency Constraints for STFT Spectrograms and Their Application to Phase Reconstruction. In Proceedings of the ITRW on Statistical and Perceptual Audio Processing, SAPA 2008, Brisbane, Australia, 21 September 2008; pp. 23–28.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.