THE COOPER UNION
ALBERT NERKEN SCHOOL OF ENGINEERING

A Partitioned Autoencoder for Audio De-Noising

by

Ethan Lusterman

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Engineering

September 2016

Professor Sam Keene, Advisor

THE COOPER UNION FOR THE
ADVANCEMENT OF SCIENCE AND ART

ALBERT NERKEN SCHOOL OF ENGINEERING

This thesis was prepared under the direction of the Candidate's Thesis Advisor and has received approval. It was submitted to the Dean of the School of Engineering and the full Faculty, and was approved as partial fulfillment of the requirements for the degree of Master of Engineering.

Dean, School of Engineering Date

Prof. Sam Keene, Thesis Advisor Date

Acknowledgements

This thesis would not be possible without the guidance and support from my advisor, Dr. Sam Keene. He has mentored me since I was an undergraduate, and I am grateful for him helping this project come to life. I also want to thank Christopher Curro, my informal second advisor who helped me to think outside the box and for whom the overall system architecture is named after.

I would like to thank Kate Thorsen for pushing me past my potential and encouraging me to stay positive despite the frustrations of research. Lastly, I would like to thank my friends and family for their support. This thesis would not have been possible without all their support.

Abstract

Traditional audio denoising systems are often linear time-invariant (LTI) and often require access to clean data to properly train to remove noise. Since clean audio is often unavailable, we build on a partitioned denoising autoencoder for denoising audio signals when clean examples are unavailable for training. In addition, the nonlinearity of a neural network architecture provides additional gains over standard linear models. We compare existing semi-supervised denoising systems as well as canonical supervised denoising autoencoders. We show that for moderate levels of noise, our autoencoder can outperform existing schemes.

Contents

List of Figures

Table of Nomenclature

1 Introduction

Advances in smartphone technology have led to smaller devices with more powerful audio hardware, allowing for common consumers to make higher quality recordings. However, recorded speech and music are subject to noisy conditions, often hampering intelligibility and listenability. The goal of denoising audio recordings is to improve intelligibility and perceived quality. A variety of applications of audio denoising exist, including listening to a recording of a band or an artist's live performance in a noisy crowd, or listening to a recorded conversation or speech under noisy conditions.

A common technique for denoising involves the use of autoencoder neural networks. [**?**] Advances in parallel graphics processing units (GPU) and in machine learning algorithms have allowed for training deeper networks faster, utilizing more hidden layers with more neurons.

Prior work in denoising audio has involved the use of noise-free training data. Since common consumers do not often have access to clean audio, we seek to denoise without the use of clean audio. Other work has touched on such a semi-supervised scenario but was used more as a preprocessing step to a classification algorithm than as time-domain denoising. [**?**]

In this thesis, we compare several neural network architectures and problem scenarios, ranging from data input types, level of noise, depth of network, training objectives, and more. In Chapter 2, we present background information on machine learning, neural networks, and signal processing as well as prior work in audio denoising. In Chapter 3, we detail the problem formally as well as introduce our signal model and sourced data. In Chapter 4, we detail all considered network architectures. In Chapter 5, we compare results

from different data inputs, levels of noise, network architectures, and training objectives and discuss methods of evaluation. Finally, we make conclusions and recommendations for future work in Chapter 6.

2 Background

2.1 Machine Learning

Machine learning involves the use of computer algorithms to make decisions based on training data. Generally, this falls into categorizing input data (classification) or determining a mathetmatical function to determine a continuous output given an input (regression). Popular classification examples include recognizing handwritten digits (MNIST) as well as determining whether an image contains a cat or a dog. (REF) An example of a regression problem is determining the temperature given a set of input features (humidity, latitude, longitude, date, etc.).

Problems where training data contain input data vectors as well as the correct output vectors (targets) are known as supervised learning problems. Training a model to denoise audio where noise was introduced to the clean audio would be a supervised learning problem. On the other hand, training a model to denoise audio where the underlying clean signal is not known is an unsupervised learning problem. Different loss (objective) functions and neural network architectures can be exploited to accomplish denoising without the clean data.

For the purposes of this thesis, we use machine learning to determine an underlying nonlinear function that removes noise from time slices of audio (i.e. regression). These slices can then be pieced back together through overlap-add resynthesis. To clarify, this is a general linear model that maps an input noisy audio vector $y[n] = x[n] + N[n]$ to $\tilde{x}[n]$, a target denoised audio vector, where $x[n]$ is the underlying clean signal and $N[n]$ is the additive background noise.

2.1.1 Regression

A classical regression technique is linear regression, where one or more independent variables x_i are used to determine a scalar dependent variable y. The case of a single independent variable x is known as simple linear regression. More formally, for k independent variables, we would like to determine a weight vector $\mathbf{w}$ and bias vector $\mathbf{b}$:

$$y_i = w_1 x_{i1} + \cdots + w_k x_{ik} + b_i, \qquad i = 1 \ldots, n \tag{1}$$

$$\mathbf{y} = \mathbf{x^T w} + \mathbf{b} \tag{2}$$

where the rows of x^T are the example input observations and $\mathbf{y}$ and $\mathbf{b}$ are column vectors.

By extension, the case of linearly estimating a vector output giving a vector input is known as a generalized linear model. A canonical example would be estimating a sine wave $x[n]$ over some number of N samples given noisy samples $y[n] = x[n] + N[n]$.

2.1.2 Overfitting and Curse of Dimensionality

2.1.3 Loss functions and Regularization

2.1.4 Gradient Stuff?

2.2 Neural Networks

In this thesis, we deal only with feed-forward neural networks, which are essentially directed acyclic graphs (DAG) for computation. In other words, information only moves through the network in one direction. An example neural network is shown in Figure 1.

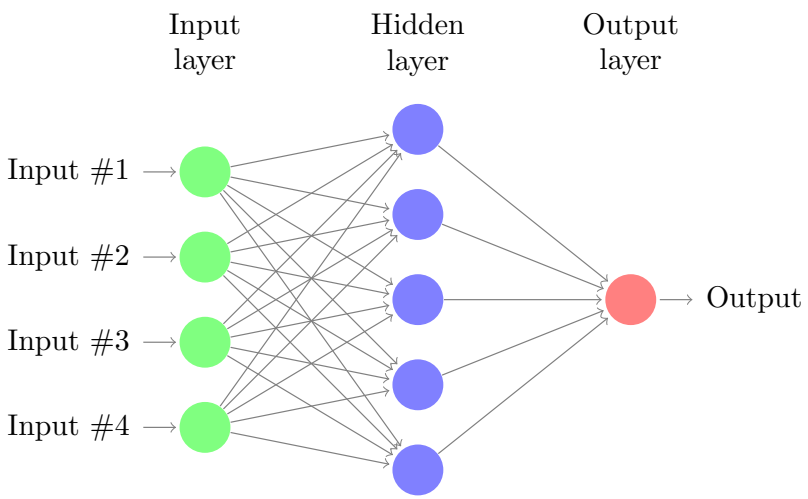

Figure 1: An example neural network. There are 4 input variables, 1 hidden layer with 5 neurons, and 1 output variable.

The connections in a nerual network can be represented by linear combinations of the input variables with learned weights $\mathbf{w}$. [**?**] Unlike standard linear models however, neural networks apply a nonlinear activation $f(\cdot)$ at the output of each neuron. The circle nodes in a neural network diagram can be thought of as the sum of the linear combinations of the connection edges and the application of the bias and activation function. Therefore, a hidden neuron z_j in a network with N input variable nodes, M hidden nodes, and K output nodes takes on the value

$$z_j = f(a_j) \tag{3}$$

where the activiation a_j is given by

$$a_j = \sum_{i=1}^{N} w_{ji}^{(1)} x_i + w_{j0}^{(1)} \tag{4}$$

The connection values w_{ji} are referred to as weights, and the scalars w_{j0} are referred to as biases. Note that the superscripted numbers refer to the Then, the output y_k is given by

$$y_k = g(a_k) \tag{5}$$

where the output activation ak is given by

$$a_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + w_{k0}^{(2)} \tag{6}$$

We are free to choose activation functions, which we will discuss later. However, note that at the output, the function $g(\centerdot)$ is often an identity for regression problems and a sigmoid $\sigma(\centerdot)$ for classification problems.

Often, the weights and biases are grouped into a weight vector $\mathbf{w}$. In other words, similar to the linear models described earlier, a neural network is a nonlinear function of input variables $\{x_i\}$ to output variables $\{y_k\}$ where the parameters of the function are learned via training techniques.

2.2.1 Dense Layer

Described in the previous section, we refer to a dense layer as a fully connected neural network, in which no interconnections between neurons are missing at each layer. Dense layers can be prone to overfitting. However, as we mention later, overfitting is not an immediate concern for the purposes of this thesis.

2.2.2 Autoencoder

An autoencoder is

2.2.3 Convolutional Layer

2.2.4 Nonlinearity Choice

2.2.5 Minibatches

2.2.6 Batch Normalization

2.2.7 Weight Updates

2.3 Signals and Systems

Domain knowledge of discrete audio signals and systems better informs our decisions for an audio denoising system, so some background information on signals and systems as it pertains to this thesis is detailed below.

2.3.1 Signals

We deal exclusively with discrete-time audio signals in this thesis. A discrete-time audio signal $x[n]$ is represented as a sequence of numbers (samples), where each integer-valued slot n in the sequence corresponds to a unit of time based on the sampling frequency f_s. This comes from sampling the continuous-time audio signal $x_c(t)$:

$$x[n] = x_c(nT) \tag{7}$$

where $T = 1/f_s$. For example, a 1-second speech signal sampled at 8kHz has 8000 samples. Furthermore, digital signals also have discrete valued sample amplitudes. For the purposes of this thesis, the bit depths of computers we use for analysis are high enough to allow for perfect reconstruction between continuous-time signals and digital signals.

We also assume signals collected have been properly sampled according to the Nyquist-Shannon sampling theorem, which states that a discrete-time signal

must be sampled at at least twice the highest frequency present in the signal to prevent aliasing of different frequencies. For example, speech signals genearlly have information up to 8kHz, so many speech signals are sampled at 16kHz. Music is more complex in that signals often span up to about 20kHz, so CD quality recordings are often sampled at 44.1kHz or higher. For this thesis, we use recordings sampled at 44.1kHz or lower.

2.3.2 Convolution

The discrete-time convolution operation takes two sequences $x[n]$ and $h[n]$ and outputs a third sequence $y[n] = x[n] * h[n]$:

$$y[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k] \tag{8}$$

Convolution is commutative, so $x[n] * h[n] = h[n] * x[n]$ holds true.

A linear, time-invariant (LTI) system is characterized by its impulse response $h[n]$, which allows us to determine samples $y[n]$ when $x[n]$ is subject to $h[n]$. For the purposes of this thesis, our underlying clean signal $x[n]$ might be subject to the conditions of an acoustic environment $h[n]$ and crowd noise $N[n]$:

$$y[n] = h[n] * x[n] + N[n] \tag{9}$$

In this scenario, our system would attempt to recover $h[n] * x[n]$ and possibly even $x[n]$ if the acoustic environment were deemed "noisy enough" due to echo and reverberation.

One of our proposed systems also incorporates convolutional neural networks (CNN) which use convolutions between frames of samples instead of simple linear combinations (discussed later).

2.3.3 Frequency Transforms

In some of our proposed systems, we use a frequency transformed version of the input signal as a preprocessing step to the system input. While no new information is gained from transforming the input, networks often respond better to determining the value of the magnitude of varying frequencies at a time slice instead of the individual time samples.

The frequency transform we use in this thesis is the discrete-time Fourier transform (DTFT). A sequence of N discrete-time samples is transformed into another sequence of N samples where each index then corresponds to a frequency bin. The DTFT $X[k]$ of a signal $x[n]$ is given by the following:

$$X[k] = \sum_{n=0}^{N-1} x[n] W_N^{kn} \tag{10}$$

where the twiddle factor W_N is given by $W_N = e^{-j(2\pi/N)}$. Then the reconstruction of $x[n]$ from $X[k]$ is given by:

$$x[n] = \frac{1}{N} \sum_{k=0}^{N-1} X[k] W_N^{-kn} \tag{11}$$

In this thesis, we also exploit the main duality between the time and frequency domain using the convolution theorem, which states that convolution in time is equivalent to multiplication in frequency and vise versa:

$$\mathscr{F}\{h[n] * x[n]\} = H[k]X[k] \tag{12}$$

$$\mathscr{F}^{-1}\{H[k] * X[k]\} = h[n]x[n] \tag{13}$$

This allows us to effectively treat our network as a non-linear filter that can denoise small time/frequency slices of our noisy signal, which can then be pieced back together using overlap-add resynthesis. We detail this in the next section.

2.3.4 Windowing and Perfect Reconstruction

To window a signal is to multiply a window function $w[n]$ by the frame, i.e. $w[n]x[n]$ over the frame length N. Because we are training a network to denoise small segments of a larger audio signal, we window the signal segments. This accommodates the finite-length requirement of the DTFT and helps to prevent spectral leakage. [DSPBOOK]

Also, to be able to properly reconstruct our signal, we use a window function and corresponding overlapping frame percentage to accomplish perfect reconstruction. The corresponding overlapping frame percentage is set such that the window sums to a constant for all time. For example, a rectangular window $w[n] = 1$ over an interval of length N has an overlap of 0% to sum to a constant 1 for all time. Another popular window is the Hanning window, defined over an interval N by the following:

$$w[n] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right) \tag{14}$$

For the Hann window, the perfect reconstruction overlap is a frame length of $N = 50\%$.

2.3.5 Window Size and Frequency v. Time Resolution Tradeoff

We must consider window size as a hyperparameter to our system. In general, shorter windows give rise to better time resolution at the cost of frequency resolution. On the other hand, longer windows give rise to better frequency resolution at the cost of time resolution. To illustrate, consider FIGURE.

2.3.6 Noise and Signal-to-Noise Ratio

Since we are trying to denoise audio signals, we must discuss how we measure noise. One of the most common measures of degradation of signal quality from additive noise is signal-to-noise ratio (SNR), defined as the ratio of signal variance to noise variance. [DSP] For the signal $y[n] = x[n] + N[n]$, where $x[n]$ is the signal of interest and $N[n]$ is the additive noise, the SNR is defined as

$$SNR = \frac{\sigma_x^2}{\sigma_n^2} \tag{15}$$

where σ^2 refers to the variance of the signal in question over some time interval. For the purposes of this thesis, we achieve desired a desired SNR for a simulation by scaling the noise to match the variance to the signal, then scaling the noise or the signal to achieve the desired SNR.

2.3.7 Magnitude and Phase Spectrum

3 Signal Model and Data

3.1 Network Input and Output

To simulate an audio denoising scheme, we define the following inputs and outputs. We take a known clean signal $x[n]$ which we subject to additive noise $N[n]$ using a specified SNR, resulting in the following noisy signal $y[n]$:

$$y[n] = x[n] + N[n] \tag{16}$$

To achieve a particular average SNR per simulation, we take the average signal energy for each minibatch of size B to determine a multiplicative scale factor k on the noise signal $N[n]$. For example, for additive white Gaussian noise (AWGN), we sample from the zero-mean, unit variance normal distribution ("randn" in Python) and determine our scale factor k as σ using the specified SNR in decibels:

$$\sigma_n^2 = \frac{1}{SNR_{lin}} \frac{1}{BN} \sum_{b=0}^{B-1} \sum_{n=0}^{N-1} x_b^2[n] \tag{17}$$

where SNR_{lin} is given by

$$SNR_{lin} = 10^{\frac{SNR_{db}}{10}} \tag{18}$$

In supervised scenarios, we allow the network to train with access to the ground truth $x[n]$. On the other hand, in semi-supervised scenarios, we only allow the network to train with access to a "soft label" indicating if the signal is (1) noise-only or (2) noise and possibly signal. [DanStowell] However, in both supervised and semi-supervised scenarios, our neural network input is one of the following:

1. Frames of $y[n]$

2. Frames of $\|Y[k]\|$

3. Magnitude spectroram frames of $Y[k]$

4. Complex spectrogram frames of $Y[k]$

We choose the frame length L, time-domain window $w[n]$, and frame overlap percentage p as hyperparameters. Generally, we use 1024-sample frames at 16 kHz with a Hanning window with 50% overlap unless otherwise specified. In addition, for frequency frames, we use an FFT length the same length as our frame for a total of $L/2$ frequency bins. Note that our choice of frame length and sampling rate allows us to balance time and frequency resolution. With the given frame length and sampling rate, we achieve a frequency resolution of 15.625 Hz/bin by the following:

$$\frac{f_s/2\,\text{Hz}}{N/2\,\text{bins}} = \frac{fs}{N} \tag{19}$$

$$= 15.625\,\text{Hz/bin} \tag{20}$$

Similary, our time resolution is given by

$$\frac{N}{fs} = 64\,\text{msec} \tag{21}$$

Since we want to evaluate the level of denoising in the time domain, we recombine the network outputs with the noisy phase components of the spectrum if necessary to obtain an estimate $\hat{x}[n]$. We then compare $\hat{x}[n]$ to $x[n]$, in general using the mean squared error (MSE). For example, when our network outputs frames of $\|\hat{X}[k]\|$, we take the inverse Fast Fourier transform (IFFT) using the

noisy phase $\angle Y[k]$ and use overlap-add to recombine the frames. (Anything to add about phase denoising and failures here?)

3.2 Signal and Noise Choices

Our choice of signals include the following:

1. Sine waves with multiple frequencies and random amplitudes and phases
2. Clean speech signals
3. Studio music recordings
4. Live concert recordings

Similarly, our choice of noise signals include the following:

1. Additive white Gaussian noise (AWGN)
2. Restaurant noise

As mentioned above, we can use the average energy per minibatch to specify a given SNR for an experiment. We take several combinations of clean and noise signals and compare across multiple SNRs.

3.3 Other Network Parameters

Since our networks involve one or more neural network layers, we show some results compared to choices of nonlinearity, number of layers (depth), and number of nodes in each layer (width). Generally, we use an identity at the network output and either the rectified linear unit (ReLU), a modified ReLU (mReLU), leaky rectify, hyperbolic tangent (tanh), or an exponential linear unit (elu).

4 De-noising Architectures

In the following sections, we detail all considered shallow network architectures. Note that these network architectures can easily be extended to deep networks by adding corresponding encode and decode layers before and after the latent representation, respectively. These networks can be trained using the various inputs detailed in Chapter 3. However, for the purposes of presenting first results, we consider single FFT frames only to compare networks.

4.1 Supervised Autoencoder

We adopt the shallow supervised autoencoder from [**?**]. Used for supervised denoising, we adopt the relative network size as well as their modified nonlinear activation function. The network structure is a single hidden layer, dense neural network. In other words, we can represent our network output $\hat{X}_i[k]$ for various overlapping frames $i = 1, \cdots, N$ by the following:

$$\hat{X}_i[k] = f_1\big(\mathbf{W}^{(1)}\mathbf{h}_i^{(0)} + \mathbf{b}^{(1)}\big) \tag{22}$$

$$\mathbf{h}_i^{(0)} = f_0\big(\mathbf{W}^{(0)}Y_i[k] + \mathbf{b}^{(0)}\big) \tag{23}$$

This network is trained to estimate the various layer weight matrices $\mathbf{W}^{(l)}$ and layer bias vectors $\mathbf{b}^{(l)}$.

Since we are estimating a magnitude spectrogram for values in the interval $[0, \infty)$, we use a nonlinear activation function whose support is on the same interval. A natural choice is the rectified linear unit (ReLU). However, as detailed in [**?**], the ReLU is subject to a 0-derivative for negative values. The modified ReLU used in [**?**], which we denote as mReLU, is given by the following: