# Contents

**Appendices** **3**

**A Time Series** **5**

A.1 Signal Processing . . . . . 5

A.2 Speech Properties . . . . . 5

**B Transforms** **7**

B.1 Fourier Transform . . . . . 7

B.1.1 Spectrums . . . . . 7

B.1.2 Discrete Fourier Transform . . . . . 7

B.1.3 Short-Time Fourier Transform . . . . . 7

B.2 Wavelet . . . . . 7

**C Spectrograms** **9**

C.1 Linear and Log . . . . . 10

C.1.1 Amplitude vs Decibels . . . . . 10

C.1.2 Mel-Bins . . . . . 10

C.2 The Phase Problem . . . . . 10

C.2.1 Linear . . . . . 10

C.2.2 Decibel . . . . . 10

C.2.3 Mel-bin . . . . . 10

C.3 Phase Retrieval Techniques . . . . . 10

C.3.1 Phase Storage . . . 10
C.3.2 Griffin-lin Algorithm . . . 10
C.3.3 Vocoders . . . 10

**D Machine Learning** **11**

D.1 Supervised vs Unsupervised Learning . . . 11
D.1.1 Supervised Learning . . . 11
D.1.2 Unsupervised Learning . . . 12
D.2 Advancements . . . 12
D.2.1 AREAS** . . . 12
D.3 Problems . . . 12
D.3.1 Data . . . 12
D.3.2 Interpretability . . . 12
D.3.3 Overfitting . . . 12
D.3.4 Hyper-parameters . . . 12
D.3.5 Computation . . . 12

**E Deep Learning** **13**

E.1 Neural Networks . . . 14
E.1.1 Structure . . . 14
E.1.2 Back-Propagation Algorithm . . . 14
E.2 Image Processing . . . 14
E.2.1 Convolutional Neural Networks . . . 14
E.2.2 Attention . . . 14
E.3 Segmentation . . . 14
E.3.1 Medicine . . . 14
E.3.2 U-Net . . . 14

# Appendices

# Appendix A

# Time Series

## A.1 Signal Processing

## A.2 Speech Properties

# Appendix B

# Transforms

## B.1 Fourier Transform

### B.1.1 Spectrums

### B.1.2 Discrete Fourier Transform

### B.1.3 Short-Time Fourier Transform

## B.2 Wavelet

# Appendix C

# Spectrograms

As we saw before, audio can be better represented as mix of different frequencies that define the sound we are hearing. The problem lies in the fact that, in speech, these frequencies change drastically during each conversation, sentence, word and syllable! This means that the perfect spectrogram would need to be a continuous representation of each moment of time and for every frequency.

This is impossible for various reasons, one being the limitations of computational representations, which forces us to make a discrete representation instead. But even if that wasn't the case, the fourier transform doesn't work on single points in time, instead it works on entire signals. That said, we can use the short time fourier transform to give us a spectrogram that is discrete in time. Combined with the discrete fourier transform, we will also have a discrete representation in the frequency axis.

Spectrograms have 3 dimensions, similar to how images are represented, but not exactly. In any given image there are two dimensions that represent space and one that represents intensity. Meanwhile Spectrograms have one dimension for time, one for frequency and the last represents the strength and sometimes the phase of the frequency signal at that time. Although some writers, the standard tends to place frequency as height, time as width and color as signal intensity in a given point.

Although this is a great step forward to giving us a better representation of audio, there are still too many parameters that play a role. Each one of these parameters affects how well the spectrogram can correctly contain the information in the audio. Although there are already well established defaults for most human tasks, it still hasn't been well studied in great part of automated and machine learning tasks.

The exception to this being Speech Recognition, but even so they haven't been heavily studied. There are few papers and datasets that can be used to prove when some parameters are better than others or if there are conservitive settings that tend to work consistently in all problems. In the end, some authors question if spectrograms are even the representation that we are looking for for various reasons that we will see in more detail later in Chapter 7.

For now we'll hide our skepticism and focus on the advantages of using spectrograms.

The spectrograms we have been referring to so far are what are called linear-spectrograms. Although these are the simplest and easiest to implement, it is not the only one. We will mention the difference between Amplitude and Decibel Spectrograms as well as Frequency and Mel-Bin Spectrograms.

## C.1 Linear and Log

The values in amplitude are represented in the "color" dimension, giving us the intensities that we see in the spectrogram image. As we saw in the Second Chapter, the amplitude tells us the strength of the signal. In the spectrogram, however, there may also be a phase value. This is because the signals tend to be represented as a Complex number, which can be expressed in cartesian or polar coordinates.

### C.1.1 Amplitude vs Decibels

### C.1.2 Mel-Bins

## C.2 The Phase Problem

### C.2.1 Linear

### C.2.2 Decibel

### C.2.3 Mel-bin

## C.3 Phase Retrieval Techniques

### C.3.1 Phase Storage

### C.3.2 Griffin-lin Algorithm

### C.3.3 Vocoders

# Appendix D

# Machine Learning

"A machine learning algorithm is an algorithm that is able to learn from data."[1] These algorithms search for patterns in the data to discover rules of association that could be used to solve the problem without explicit instructions. This approach turns out to be better, in various task, than previous hard-encoded methods. These older algorithms are usually referred to as traditional methods.

Most of the success of machine learning systems stem from the fact that it is hard to give a perfect definition of something. To achieve this we would be required to tumble into epistemological questions. Most would agree that a definition of a cat would be incomplete without mentioning a tail, yet a cat without a tail is still a cat.

## D.1 Supervised vs Unsupervised Learning

Machine Learning algorithms are commonly separated into supervised and unsupervised learning. Even though each of these methods have their pros and cons, most of the current study has focused on developing supervised methods. Part of the reason is because of how data is currently being collected in the real world.

Apart from these two branches of machine learning there is another call "semi" supervised learning[2]. This approach uses a mix of labeled and unlabeled data items. In real life this allows users to quickly add labels to a dataset. This is done with unsupervised methods, which are later "corrected" by a human user which labels the faulty items.

Even though these tools have proven useful in certain problems, including large scale dataset labeling, they are not relevant to this study.

### D.1.1 Supervised Learning

Supervised learning methods, as we mentioned before, use labeled datasets as training data. This allows the algorithm to repeatedly find patterns which could allow it to gain

better insights. The training process for these algorithms is similar to a student studying with a mock exam.

In this analogy, our mock exam is the same as our dataset. Even so, just like in real life, we elaborating a mock exam isn't always easy. How do we provide questions similar to the real test, without making it too similar? If you told a class of students that the real exam will be exactly the same as the mock, the students would be less inclined to learn the actual concepts necessary to pass the test (or any similar test). Instead, they might try to memorize the answers knowing that this would be enough.

This same login is what happens when training an algorithm. They don't have a notion of what they are trying to learn but instead what they want to answer. A program is (at least until now) unable to understand that the dataset is just a mock, and that the real test comes after. What's worse, while a normal student would have problems memorizing tests for very long, a computer store these answers with no additional punishment.

### D.1.2 Unsupervised Learning

## D.2 Advancements

### D.2.1 AREAS**

## D.3 Problems

### D.3.1 Data

### D.3.2 Interpretability

### D.3.3 Overfitting

### D.3.4 Hyper-parameters

### D.3.5 Computation

# Appendix E

# Deep Learning

## E.1 Neural Networks

### E.1.1 Structure

#### Inputs

#### Weights

#### Bias

### E.1.2 Back-Propagation Algorithm

#### Forward Propagation

#### Gradients

## E.2 Image Processing

### E.2.1 Convolutional Neural Networks

### E.2.2 Attention

## E.3 Segmentation

### E.3.1 Medicine

### E.3.2 U-Net

# Bibliography

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[2] O. Chapelle, B. Schölkopf, and A. Zien, "Semi-supervised learning explained," 2006.