Chapter 1

Introduction

1.1 The Cocktail Party Problem

Imagine being in a public area with one of your friends. As you talk about your work and what you learned in school there are people around you making noise. They're also talking about their lives, what they plan on doing in the mall or the movie that just came out. Kids laughing with their parents talking, nearby traffic with car horns flaring in the distance. If you wanted to, you could hear them, but instead you ignore it.

You continue to listen to your friend, regardless of the noise in the background. You have no problem paying attention until suddenly you hear your name being mentioned in the crowd. For some reason, somewhere, someone said your name. You look around to see a familiar face walking towards you...

What we described in the short and imaginary story above is what is called the "Cocktail Party Effect". This effect was first coined by Cherry Colin[1] where he proposed a series of tests that would measure the limits of a human's ability to listen to a specific voice under different circumstances. The problem of getting a machine to do this same task was called the Cocktail Party Problem by Cherry.

The cocktail party problem is one of the biggest unsolved problems in computation. The short and imaginary story above gives us an example of how we, as humans, solve this problem in a seemingly effortless way. Thanks to evolution, humans can it so effortless, in fact, that most people don't even stop to appreciate how complicated the task actually is.

Its difficulty, however, is not the reason it is so widely studied. In fact, this small problem seems to be the barrier that has kept us from advancing in the automation of automation, or at least in the way we would like. The cocktail party is not only present in sound, but in just about any signal processing problem, from medical scanning to telecommunications[2], noise always seems to find its way into our sensors.

That is why a general solution to this problem will not only allow us to improve greatly in audio related tasks, such as speech recognition, transcriptions, audio classifi-

cation, and audio/speech enhancements, but also in various fields like medical analysis and seismology. The reach of these advancements also promises a wide range of new technologies shortly after.

This is why, in this work, we will attempt to take a dive into how we can leverage the current improvements in artificial intelligence in tackling the problem. Ever since the latest boom of deep learning [3]

Even though, as we saw before, the problem generalizes beyond speech, we will limit our research to stick with voiced data. This is due to the complexity in examining signals as a whole, and because of the focus that most methods currently have on the subject.

Given how the difficulty of the problem lays much beyond my current limits as a researcher we will not show. We will accept limiting our problem greatly for practical reasons, in an attempt to get insight on . This also includes a review of how traditional methods attempt to solve the problem.

1.2 History

Speech separation has been a problem that researchers have been interested in for years. So much so that the problem was formulated decades ago by Colin Cherry[1]. In his famous paper he gives an example of the task with a conversation in a Cocktail party, giving the problem its name. Over 65 years later one could argue that progress is just now being made in the area, mostly due to the advancements in general learning algorithms, like deep learning, giving an edge in unstructured data analysis.

1.3 Document Structure

In this project we will separate our investigation into three parts. The first part will review the current literature on the problem, focusing specifically on current processing methods. We then explore how previous studies used these methods and the possible affects this has on the results. This will be split into 4 Chapters.

Chapter 1 will give a short rundown on the current data processing techniques. We will explain why data preprocessing is important and what problems occur without it, focusing on audio tasks. This chapter is divided into two parts, the first will touch on feature extraction for audio. The second part will touch on Normalization.

For feature extraction we will need to learn about Spectrograms. The Spectrograms can be separated into two commonly used categories: linear and mel-bin. Both spectrograms represent audio clips into a 3-D transformations, similar to images. A linear spectrogram has a linear frequency dimension while mel-bins represents frequency in logarithmic scale.

Chapter 2 will reference one of the problems we currently have in Generating Audio. Here we explain why generating audio directly isn't always possible and what the modern solution for these problems currently looks like. The critiques given to these current

solutions will be based on the literature and on my own personal experience.

Chapters 3 and 4 will talk about recent projects. Although these projects do not tackle the problem in the same way, they have reasonably good quality in audio. Because of this, we will use these studies as a reference to what practices could be used to provide these results. Later on we will see that one uses vocoders and the other changes the training target.

At the end of these chapters we will state the pros and cons of each methods and briefly show the train of thought that lead to the experiments that we can find in part 2.

Part two contains the methodology that was taken in this study. In this we hope to explain the experiment well enough to be replicated in future studies. For this we provide a context of the algorithm and how each process was implemented. As we will learn later, the implementation of some of these algorithms which varies in the different libraries affect the outputs similar to the parameters.

Finally the third and final part of this text will give the results to the project. In this we will explain what the quantified errors represent and a quick qualitative assessment of the audios. ***As a reminder, we do not plan on building a model that competes against the current SoTA. ***Instead we will attempt to find input and target creation techniques that improve the quality of the generated audio.

The final part of this thesis contains the appendix and the bibliography. In the appendix we give most of the theoretical background necessary to understand the methods used in our research. This includes subjects ranging from machine learning, signal processing, the fourier and wavelet transform and a small usage guide for the source code and where to find it.

Bibliography

[1] C. E. Colin, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] L. Marchegiani, S. Karadogan, T. Andersen, J. Larsen, and L. Hansen, "The Role of Top-Down Attention in the Cocktail Party: Revisiting Cherry's Experiment after Sixty Years," in *Proceedings of the tenth International Conference on Machine Learning and Applications (ICMLA'11)*, (United States), IEEE, 2011.

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