Exploring the Cocktail Party

A review of

Héctor M. de la Rosa Prado

A thesis presented for the degree of
Bachelor in IT in Science

Department Name
National Autonomous University of Mexico
Mexico
November 11, 2019

Contents

Thesis Title

Thesis Subtitle

Héctor M. de la Rosa Prado

Abstract

Chapter 1

Introduction

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 classification, 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

1.1 History

1.2 Acknowledgements

Chapter 2

The Cocktail Party Problem

2.1 Segmentation vs Attention Problem

2.2 Ill-Posed Problems

2.3 Inverse Problems

Chapter 3

Time Series

3.1 Signal Processing

3.2 Speech Properties

Chapter 4

Transforms

4.1 Fourier Transform

4.1.1 Spectrums

4.1.2 Discrete Fourier Transform

4.1.3 Short-Time Fourier Transform

4.2 Wavelet

Chapter 5

Spectrograms

5.1 Linear vs Log

5.1.1 Amplitude vs Decibels

5.1.2 Mel-Bins

5.2 The Phase Problem

5.2.1 Linear

5.2.2 Decibel

5.2.3 Mel-bin

5.3 Phase Retrieval Techniques

5.3.1 Phase Storage

5.3.2 Griffin-lin Algorithm

5.3.3 Vocoders

Chapter 6

Machine Learning

6.1 Supervised vs Unsupervised Learning

6.1.1 Supervised Learning

6.1.2 Unsupervised Learning

6.2 Advancements

6.2.1 AREAS**

6.3 Problems

6.3.1 Data

6.3.2 Interpretability

6.3.3 Overfitting

6.3.4 Hyper-parameters

6.3.5 Computation

Chapter 7

Deep Learning

7.1 Neural Networks

7.1.1 Structure

Inputs

Weights

Bias

7.1.2 Back-Propagation Algorithm

Forward Propagation

Gradients

7.2 Image Processing

7.2.1 Convolutional Neural Networks

7.2.2 Attention

7.3 Segmentation

7.3.1 Medicine

7.3.2 U-Net

Chapter 8

Literature Review

8.1 Current SoTA with AI

8.1.1 Speech Recognition

8.1.2 Speech Separation

Looking to Listen

8.2 Feature Extraction and Normalization

8.2.1 Do Spectrograms belong in AI?

8.3 Data Augmentation

8.4 Targets and Masks

Chapter 9

Methodology

9.1 Computer Requirements

9.2 Dataset

9.3 Preprocessing

9.4 Model

9.5 Targets

9.6 Error

Chapter 10

Implementation

10.1 Libraries

10.1.1 Librosa

10.1.2 PyTorch

10.1.3 TorchAudio

10.2 Source Code

10.2.1 Running Instructions

Chapter 11

Results

11.1 Initial Dataset

11.1.1 Quantified Losses

11.1.2 Quality Tests

11.1.3 Samples (images and audios)

11.2 Transfer learning

11.2.1 Quantified Losses

11.2.2 Quality Tests

11.2.3 Samples (images and audios)

11.3 Transfer Learning with fine tuning

11.3.1 Quantified Losses

11.3.2 Quality Tests

11.3.3 Samples (images and audios)

Chapter 12

Discussion

12.1 Future Work

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.