BIOLOGICALLY-INSPIRED AUDITORY ARTIFICIAL INTELLIGENCE FOR SPEECH RECOGNITION IN MULTI-TALKER ENVIRONMENTS

LUKAS GRASSE
Bachelor of Science, University of Lethbridge, 2018

A thesis submitted
in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in

NEUROSCIENCE

Department of Neuroscience
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

BIOLOGICALLY-INSPIRED AUDITORY ARTIFICIAL INTELLIGENCE FOR SPEECH RECOGNITION IN MULTI-TALKER ENVIRONMENTS

LUKAS GRASSE

Date of Defence: October 20, 2020

Dr. Matthew Tata Thesis Supervisor	Professor	Ph.D.
Dr. Artur Luczak Thesis Supervisor	Professor	Ph.D.
Dr. Aaron Gruber Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. John Zhang Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. David Euston Chair, Thesis Examination Committee	Associate Professor	Ph.D.

Dedication

To my brother, mom and dad, and grandma Ruth.

proverbium est adulescens iuxta viam suam etiam cum senuerit non recedet ab ea.

Abstract

Understanding speech in the presence of distracting talkers is a difficult computational problem known as the cocktail party problem. Motivated by auditory processing in the human brain, this thesis developed a neural network to isolate the speech of a single talker given binaural input containing a target talker and multiple distractors. In this research the network is called a Binaural Speaker Isolation FFTNet or BSINet for short. To compare the performance of BSINet to human participant performance on recognizing the target talker's speech with a varying number of distractors, a "cocktail party" dataset was designed and made available online. This dataset also enables the comparison of network performance to human participant performance. Using the Word-Error-Rate metric for evaluation, this research finds that BSINet performs comparably to the human participants. Thus BSINet provides significant advancement for solving the challenging cocktail party problem.

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my co-supervisor Dr. Matthew Tata for his continuous guidance and support throughout my master's degree. The combination of guidance and freedom to explore new projects and ideas in his lab has made graduate school an amazing and enjoyable experience. I would also like to express thanks to my co-supervisor Dr. Artur Luczak for his guidance and advice throughout my degree.

Next, I would like to thank my committee members Dr. Aaron Gruber and Dr. John Zhang. Their valuable feedback and insight during committee meetings helped keep me on track. I would also like to thank Naomi Cramer for all her valuable help with the administrative side of being a graduate student and dealing with different offices and departments in the university. I would also like to thank Gene for keeping me hydrated over the years.

I am also thankful to Dr. Francesco Rea for hosting me during my research period at IIT, Jonas Gonzalez for his collaborations and guidance, and all the other members of the RBCS lab for their hospitality.

Finally, I express sincere thanks to my friends and family for supporting and encouraging me throughout this degree.

Contents

Contents vi

List of Tables viii

List of Figures ix

1 The Auditory System and Speech in Multi-Talker Environments 1
- 1.1 Preface . . . 2
- 1.2 Spatial Hearing in the Mammalian Brain . . . 4
  - 1.2.1 The Early Auditory System . . . 4
  - 1.2.2 The Characteristics of Spatial Auditory Cues . . . 6
- 1.3 Summary . . . 8

2 Binaural Speaker Isolation Neural Networks 9
- 2.1 Introduction . . . 9
  - 2.1.1 Traditional Computational Approaches . . . 9
  - 2.1.2 State of the Art Computational Approaches . . . 12
- 2.2 Methodology . . . 14
  - 2.2.1 Dataset Preprocessing . . . 14
  - 2.2.2 System Architecture . . . 17
  - 2.2.3 Training Process . . . 20
- 2.3 Results . . . 21
  - 2.3.1 Network Experimental Setup . . . 21
  - 2.3.2 WER Evaluation . . . 22
  - 2.3.3 SDR Evaluation . . . 22
  - 2.3.4 WER Evaluation on Different Cues . . . 23
- 2.4 Summary . . . 23

3 Comparison of Network and Human Performance Across Cues 25
- 3.1 Introduction . . . 25
  - 3.1.1 Comparing Neural Networks and the Mammalian Brain . . . 26
- 3.2 Methodology . . . 28
  - 3.2.1 Human Participant Experimental Setup . . . 28
- 3.3 Results . . . 29
  - 3.3.1 WER Evaluation on HRIR Features . . . 29
  - 3.3.2 WER Evaluation on Different Cues . . . 30
- 3.4 Summary . . . 30

4 Discussion and Future Directions **32**
4.1 Discussion 32
4.1.1 Comparison with Biological Systems 34
4.1.2 Comparison with Previous Computational Approaches 36
4.2 Future Work and Applications 37
4.2.1 Architecture Improvements 37
4.2.2 Future Experiments 38
4.2.3 Applications 39
4.3 Conclusion 39

Bibliography **41**

A Appendix: Cocktail Party Speech Dataset **46**
A.1 Dataset Overview 46
A.2 Data Sources and Licensing 46

List of Tables

2.1 **SDR Improvement for Monaural and Binaural Network** 23

List of Figures

1.1 **ITD and ILD cues.** This figure shows a top-down view of a human listener hearing a sound source in the right hemifield. The top half of the figure illustrates the time delay that occurs due to the fact that the signal takes longer to reach the listener's left ear. The bottom half illustrates the level or intensity difference caused by the attenuation of the signal as it propagates through the head. 5

2.1 **BSINet training process.** The training process starts with rendering the target talker at 0°, and distracting talkers rendered randomly to the left at -90°, -60°, or -30°and to the right at 30°, 60°, or 90°. The speaker isolation network was then trained to isolate a monaural channel containing the target talker's speech. For evaluation, the target and distracting talkers were spatially rendered, then inference was performed using the speaker isolation network, and the resulting isolated audio was used as input for a speech recognition system. 15

2.2 **Original FFTNet architecture** due to [26]. The operations inside each layer are as follows: a given input is first divided into two halves; each half is then passed through a different 1x1 convolution layer and then summed together. This summed output is passed through a ReLU function, another 1x1 convolution, and a final ReLU. 18

2.3 **Binaural Speaker Isolation Network (BSINet).** The network consists of two separate layers for the left and right channels, the output of the left and right layers are then combined together and fed into a third set of layers. The network outputs monaural samples of audio for the target talker. . . . 18

2.4 **Word Error Rate (WER) performance for different networks.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of approach with F = **83.55** and a main effect of number of distractors with F = **117.18**. 22

2.5 **WER performance for BSINet with different spatial cues.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of feature type with F = **42.89** and a main effect of number of distractors with F = **148.32**. 23

3.1 **Performance comparison between human participants and BSINet.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of approach with F = **61.44** and a main effect of number of distractors with F = **141.60**. 29

3.2 **Human participant performance with different spatial cues.** The coloured bands show the 95% confidence interval of the results across all participants. A two-way ANOVA showed a main effect of feature type with F = **30.85** and a main effect of number of distractors with F = **63.85**. 30

Chapter 1

The Auditory System and Speech in Multi-Talker Environments

1.1 Preface

Humans have the remarkable ability to understand speech in noisy environments with many distracting talkers. This ability has been called the cocktail party effect, a term first coined in a paper published by E. C. Cherry in 1953 [8], or more recently the ”multi-talker problem”. Understanding speech in these multi-talker environments is challenging and the mechanisms by which the brain performs this task are still not fully understood. One clue, which emerges from a large body of previous research is that spatial hearing in general and spatial release from masking in particular are important contributors to the brain's ability to perform this task. [22][64][20][28][32][1].

Spatial hearing is the ability to discern and use the spatial configuration of sounds in the auditory scene around a listener. Sound localization, which is the ability to know the direction and distance of a sound source, is perhaps the most salient aspect of spatial hearing, and has certainly received the most attention in the literature. However, spatial hearing also provides an important dimension (together with frequency) over which the brain can resolve the complex auditory scene and solve the multi-talker problem. To achieve spatial hearing, humans and other mammals rely mainly (although not exclusively) on the binaural comparison of the auditory signals present at each ear. There are two binaural cues the brain uses to localize sounds in space: the time delay between the two ears, commonly referred to as interaural time delay (ITD), and the level difference between the two ears which is referred to as the interaural level difference (ILD). It is well known that these cues enable the brain to localize the direction of arrival of sounds. These cues also provide the putative spatial dimensions along which a target sound can be spatially separated from distracting or masking sounds that arrive from different spatial locations. Although much research has demonstrated the importance of ITD and ILD in computing direction of arrival in birds e.g. [6] and mammals e.g. [14] (reviewed e.g. in [15]), the computational mechanisms that presumably use ITD and/or ILD to perform spatial unmixing have been less investigated and remain unclear. Some previous studies have suggested that both ILD and ITD cues

are independently sufficient to mediate the benefits of spatial hearing in complex auditory scenes [11][28][13] (see discussion). The goal of this thesis was (1) to consider whether an artificial neural network could be trained to make use of these binaural cues to improve speech recognition in complex auditory scenes, and (2) to consider which, if any, binaural cues the network makes use of.

This thesis is organized in the following way. Chapter one first broadly outlines the biological systems in the brain that are involved in spatial hearing - these are the systems that compute and represent the spatial relationships between temporally overlapping sound sources. Next, chapter two reviews the traditional approaches taken in the past for computational auditory scene analysis, and then describes state-of-the-art neural network approaches. Chapter two also describes an experiment that tests the prediction that providing binaural cues to a denoising neural network can improve speech recognition of a target talker. Next, chapter three compares the results of the neural network to the results of human participants performing the same listening task. Finally, chapter four contains a discussion of the research outlined in the thesis and also explores potential future directions for the research.

1.2 Spatial Hearing in the Mammalian Brain

The research in this thesis is based on insights from biological auditory systems. The following section outlines the pathways of the early auditory system that are related to spatial hearing as well as the fundamental spatial cues the auditory system uses.

1.2.1 The Early Auditory System

The auditory system in the mammalian brain starts with the basilar membrane in the cochlea. Sound waves entering the ear transfer mechanical sound energy to the basilar membrane. This energy is transduced by hair cells on the basilar membrane which propagate signals centrally through the auditory nerve. Bushy cells in the ventral cochlear nucleus combine inputs from multiple hair cells and route the combined signals to several nuclei of the central auditory system. [30]

Along the ascending auditory pathway, the superior olivary complex (SOC) performs critical spatial computations, with specialization in different sub-regions for ITD and ILD cues: cells in the Lateral Superior Olive (LSO) are sensitive to level differences between the two ears, whereas cells in the Medial Superior Olive (MSO) are sensitive to time delays in signals between the two ears [15]. Figure 1.1 illustrates how the ITD and ILD cues are created due to the distance between the two ears and the attenuation of the signal due to the head.

Comparison of level differences between the ears happens on timescales of tens or hundreds of milliseconds, which are long enough to allow the LSO to directly combine inputs from both ears. By contrast, ITD cues require temporal resolution on the scale of microseconds. This means that the systems involved in processing ITD cues are complex, requiring specialized neural circuitry. In 1948, Lloyd Jeffress outlined a model [25] (see also [31]) that can accomplish this, in which banks of neurons from both ears are organized into delay lines that propagate the auditory signal from each ear towards each other. A bank of ”coincidence detector” neurons are connected to the delay lines. Sounds propagating from

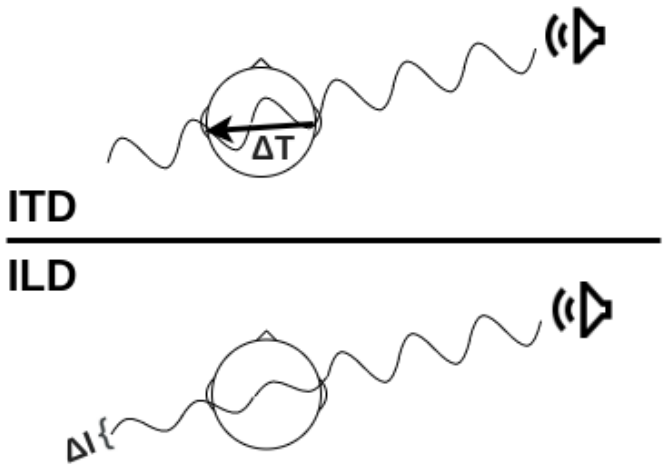

Figure 1.1: **ITD and ILD cues.** This figure shows a top-down view of a human listener hearing a sound source in the right hemifield. The top half of the figure illustrates the time delay that occurs due to the fact that the signal takes longer to reach the listener's left ear. The bottom half illustrates the level or intensity difference caused by the attenuation of the signal as it propagates through the head.

sources off of the azimuthal midline must travel different distances to the two ears and thus arrive with slight time lags relative to each other. Sensitivity of the coincidence detectors to this temporal lag imparts a degree of spatial tuning and the bank of coincidence detector neurons thus can be thought of as a place-code map of azimuthal auditory space, with an egocentric coordinate system [17], and with neurons spatially tuned to specific delays that correspond to specific azimuthal angles. This coincidence detector system has been shown to exist in the avian auditory system [65][6]. It is interesting to note, however that the common ancestor of mammals and birds did not have a tympanic ear, and as such the functional anatomy and resultant computational architecture of the auditory pathways evolved independently from each other. As a consequence, the systems responsible for processing ITD cues operate somewhat differently in the mammalian brain and are less understood than their avian counterparts. Thus the mechanism by which the ascending auditory pathway uses ITD to represent sound arrival angle is more complex than a simple system of delay lines described by the Jeffress model [15]. From a computational perspective, this also

means that computationally analogous approaches inspired by the Jeffress model, such as beamforming (see chapter 2), may not be performing the same operations as the mammalian brain.

1.2.2 The Characteristics of Spatial Auditory Cues

The brain also uses different cues for processing sounds at different frequencies and locations. An early insight into ILD and ITD cues and how they affect perception was first published by Lord Rayleigh in 1907 [43]. He discovered that the brain relies on the ITD cue for sounds at low frequencies, and as frequency increases the brain switches to using the ILD cue. This has been named the duplex theory of sound localization. The differentiation is presumed to arise because the head better attenuates sounds at higher frequencies due to an acoustic shadow, while lower frequencies are attenuated less. In concert with this, the shorter wavelengths of higher-frequency sounds also lead to spatial aliasing and thus more than one potential angle for a sound source. For example, if we picture a high-frequency sine wave with a wavelength that is much shorter than the distance between the ears, then there are multiple physiologically plausible time delays for which this sine wave would be coherent with itself at the two ears and thus at a coincidence detector in the auditory pathway. This coincidence therefore corresponds to different potential angles of which only one is correct. The combination of these two effects provide the mechanistic basis for duplex theory.

Duplex theory explains an essential component of binaural sound localization: that the mechanisms of spatial hearing interact with the frequency components of the sound source. Thus, it follows that the considerations regarding duplex theory are also important for the exploration of spatial unmixing in the cocktail party problem. This theory has led to the approach taken in papers such as [28], in which audio is filtered based on the frequency ranges described by duplex theory, with the ITD cue low-pass filtered at 1500 Hz and the ILD cue band-pass filtered at 3000-6000 Hz. While this is a valid and common approach,

the brain can use ILD cues for low frequency sounds that are near-field and can also make use of ITD cues based on the envelope of complex sounds at higher frequencies, as outlined in [15]. For this reason, the research in this thesis takes the approach of rendering ITD and ILD cues separately, and presenting the stimulus using headphones to mitigate the considerations of ITD at high frequencies and ILD at low frequencies. This approach also has the added benefit of ensuring speech signal degradation occurs only from distracting talkers and not from the effect of filtering out of potentially important frequencies by the head.

1.3 Summary

In summary, this introduction started out by exploring the pathways of the early auditory system that are involved in spatial hearing. We then considered the spatial auditory cues that are vital to the auditory system's ability to localize sounds and spatially unmix complex auditory scenes with multiple talkers. An area of the ascending auditory pathway known as the superior olivary complex (SOC) was outlined as a site where critical spatial computations are performed. These computations involve the use of two important binaural cues: the interaural time delay (ITD) and interaural level difference (ILD) cues. An important theory that shows how the brain uses ITD and ILD cues at different frequencies is duplex theory. Duplex theory shows how the brain relies on the ITD cue for localizing sounds at low frequencies, and as frequency increases the brain switches to using the ILD cue.

The next chapter will investigate traditional and current computational approaches to unmixing speech in multi-talker environments and then demonstrate the computational approach taken in this thesis: a type of neural network inspired by binaural hearing in birds and mammals called a Binaural Speaker Isolation FFTNet or BSINet for short.

Chapter 2

Binaural Speaker Isolation Neural Networks

2.1 Introduction

Although humans are adept at solving the cocktail party problem, it is still a difficult problem to model computationally and implement in artificial hearing systems that use microphones. This chapter will first take a look at traditional computational approaches and then state-of-the-art approaches based on neural networks. A common technique used in solving this problem is to make use of arrays of many microphones in order to increase performance. Increases in the number of microphones generally correspond to increased ability to resolve target talkers in these complex environments. We know from the previous chapter, however, that humans have the remarkable ability to solve this problem using only two ears. Based on this insight, the purpose of this thesis is to test the theory that binaural input can provide the computational basis for unmixing speech in multi-talker environments. To that end, the second part of this chapter will introduce the computational approach taken in this thesis, a neural network called a Binaural Speaker Isolation FFTNet or BSINet for short. This chapter also contains an experiment that tests the extent to which the BSINet has learned to use the ITD and ILD cues.

2.1.1 Traditional Computational Approaches

There have been various approaches taken to isolate a target talker from distractors, and they generally can be categorized into single-channel or multi-channel approaches. Single-

channel approaches have been based on a field of psychoacoustic research called auditory scene analysis [4]. Computational solutions based on this research such as [10][23] are called computational auditory scene analysis and typically use filterbanks to convert the audio signal into a time by frequency (T-F) representation. Once the audio is in this representation, different T-F units are grouped together based on psychoacoustic features of the target talker such as the pitch of the talker's voice and/or temporal correlations [52][58][47].

Single-channel computational auditory scene analysis approaches are interesting due to similarities with the initial frequency decomposition performed by the basilar membrane of the mammalian auditory system, however monaural signals do not provide the spatial cues that can be highly effective in unmasking a target talker from distractors. By contrast, multi-channel systems that can spatially isolate the target talker have been developed. These fall broadly into two computational approaches: Beamforming, based on relative phase lags of signals at spatially separated microphones in an array, and blind source separation such as Independent-Components Analysis, based on relative level differences of component signals in the mixtures at each microphone. Beamforming and blind source separation are interesting because they might be broadly analogous to biological computational mechanisms that make use of ITD and ILD, respectively, and thus may provide the basis of computational models that investigate how ITD and ILD cues are used to compute spatial unmixing in the biological auditory system.

Beamforming amplifies signals arriving from specific angle of arrival while rejecting signals arriving from other directions. This works due to the fact that signals take different amounts of time to arrive at sensors with different spatial locations. For this reason, beamforming can be thought of as similar to the ITD cue that the brain uses for sound localization and, presumably, spatial release from masking. A common type of beamforming that is widely used is delay-and-sum beamforming. Given a spatial direction from which to isolate a source signal, delay-and-sum beamformers calculate the corresponding time delay for each sensor and shift the incoming signal by the delay amount. These shifted

signals are then summed together so that the target signal is boosted. Because distracting signals from other locations are forced out of phase, they are attenuated in the superposed signal. Filter-and-sum or ”narrow band” beamforming is a similar type of beamforming that applies a filter bank to each sensor's signal before the delay-and-sum step. This helps compensate for the fact that the beam width changes based on the frequency and spacing of microphones in traditional delay-and-sum beamforming. In general, beamforming of low-frequency signals yields spatially wide ”beams” that only slightly attenuate off-beam distractors. By contrast, beamforming of high-frequency signals yields a sharply focused beam, however if the spacing of the microphones is greater than 1/2 the wavelength of the target frequency then spatial aliasing will result in ”phantom” arrival angles. Nevertheless, a beamforming-like computation based on ITD might be specifically useful for speech because speech signals are broadband. Hambrook et al. [17] described a computational basis for spatial selective auditory attention based on a beamforming mechanism, which used head rotations and a Bayesian updating approach to mitigate spatial aliasing at high frequencies.

The second type of multi-channel systems for the cocktail party problem are blind source separation systems (BSS). These systems exploit the fact that speech signals from different talkers are statistically independent in the mixed input at the two ears. Generally, these systems work by finding an unmixing matrix that optimally transforms the superposed signals at the two ears into independent components. Thus, a common technique for blind source separation is Independent Component Analysis (ICA) [24]. ICA works by starting with the assumption that each target talker signal is non-gaussian and statistically independent from other talkers and distracting sound sources. The ICA algorithm then estimates a mixing matrix that defines how the target signal could be produced by linearly mixing ”latent variable” input sources. Each input source can then be estimated using the inverse of the mixing matrix. ICA can perform well in ideal scenarios but has disadvantages in real-world settings. For example, ICA typically fails if strong reverberation is present. Fur-

thermore, given n component signals in the mixture, ICA requires at least $n+1$ sensors to unmix the sources plus noise. Consequently, ICA-based systems require many microphones to unmix all the talkers in a real-world scenario. Blind source separation approaches have an advantage in that the system does not need to have much prior information about the target talker, but this also means that these systems typically cannot take advantage of other information that might be useful, such as the spatial location of a target talker.

2.1.2 State of the Art Computational Approaches

Recently there have been vast improvements to a wide range of computational tasks through the rise of deep learning, a technique in which deep neural networks are trained to directly solve classification or regression problems. These networks have outperformed traditional techniques on a variety of tasks and are inspired generally by neuronal mechanisms of learning in the brain, which makes them a good choice for solving problems that the brain can easily solve. Deep learning has been applied to the cocktail party problem using a few different approaches which are outlined below.

A common approach is to convert an auditory scene comprised of several speakers into T-F units such as by a short-time fourier transform (STFT) and then to train a neural network to directly predict the STFT of the individual talkers as in [54][55] or by training a network to output a mask that is then applied to the T-F units of the target talker as in [59][12]. Both the direct Time-Frequency unit prediction and mask prediction approaches typically train the networks as a regression task, using the Mean Squared Error or similar as the loss function.

Another interesting approach is to train neural networks to project the Time-Frequency units to a high-dimensional space so that clustering can be applied to each Time-Frequency unit, with each cluster then corresponding to a talker in the scene. This approach has been used in the Deep Clustering[21] and Deep Attractor Network [34] architectures. These approaches improve on the previously described mask-based techniques in that they can

isolate a variable number of speakers.

One important factor to note is that most of these deep-learning approaches are only applied to a single-channel input and this means they cannot make use of important spatial cues such as ITD and ILD. An exception to this is [7] in which the authors first perform beamforming on multi-channel audio and then apply the Deep Attractor Network model to each beam.

Until recently, most deep learning approaches for tasks in the auditory domain used compact Time-Frequency representations of audio such as the STFT or Mel-frequency cepstral coefficients (MFCCs) as input features. This is due to the fact that these representations are much smaller than the original raw waveforms and map more easily onto audio domain tasks such as speech recognition. Recent papers such as WaveNet[37] have shown that it is possible to use dilated convolutional neural networks to directly synthesize speech. The WaveNet architecture was also used for speech denoising in [44]. Since WaveNet, architectures such as WaveRNN[27] and FFTNet[26] have shown massive improvements in the computational time required to run such models. Papers such as Wav2Letter[9] have shown that it is possible to train speech recognition networks directly on raw waveforms. One interesting aspect of the WaveNet and FFTNet architectures is that the dilated convolutional process of WaveNet resembles a wavelet analysis of the raw audio and the FFTNet architecture's process resembles the Fast Fourier Transform. Both of these spectral decompositions are broadly analogous to the tonotopic decomposition of sound by the basilar membrane of the ear.

In the following research we test the theory that binaural input provides a computational basis for unmixing the auditory scene by feeding binaural audio into a neural network that must learn to isolate a target talker using binaural cues. To do this we trained a model based on FFTNet to isolate a target talker from distractors in a multi-talker scenario. Our FFTNet speaker-isolation model takes two-channel audio as input and learns to output a single channel of audio containing only the target talker's speech. We call this network a Binaural

Speaker Isolation FFTNet or BSINet for short. The initial layers of the network process the left and right channels separately, similar to what happens in the early mammalian auditory system. A middle layer in our network then merges the initial layers from the two channels so that the network can learn to use spatial cues such as the ITD and ILD. We evaluated the network on a multi-talker dataset with 0 to 6 distracting talkers to show that it is more robust to increasing distraction than a monaural (i.e. non-spatial) network. Importantly, we investigated whether the network had learned to use either or both of the ITD or ILD cues by providing it with those cues in isolation and measuring the resulting decrement in robustness to distraction.

2.2 Methodology

Our goal was to develop a speaker isolation system as a front end for speech recognition that would use spatial cues in a manner inspired by biological hearing systems. To develop and test our system, our approach involved the creation of various auditory scenes of increasing complexity, so as to evaluate how speech recognition performance changed as a function of set size of distracting talkers. Our speaker isolation system is a binaural network based on the FFTNet architecture[26] that we call Binaural Speaker Isolation FFTNet or BSINet for short. In the next section we first describe our audio preprocessing, then the network architecture details, and finally our training process, which is also outlined in figure 2.1.

2.2.1 Dataset Preprocessing

In the following section we outline the dataset creation process. The dataset is available online[1].

[1] https://github.com/lukeinator42/cocktail_party_audio_dataset

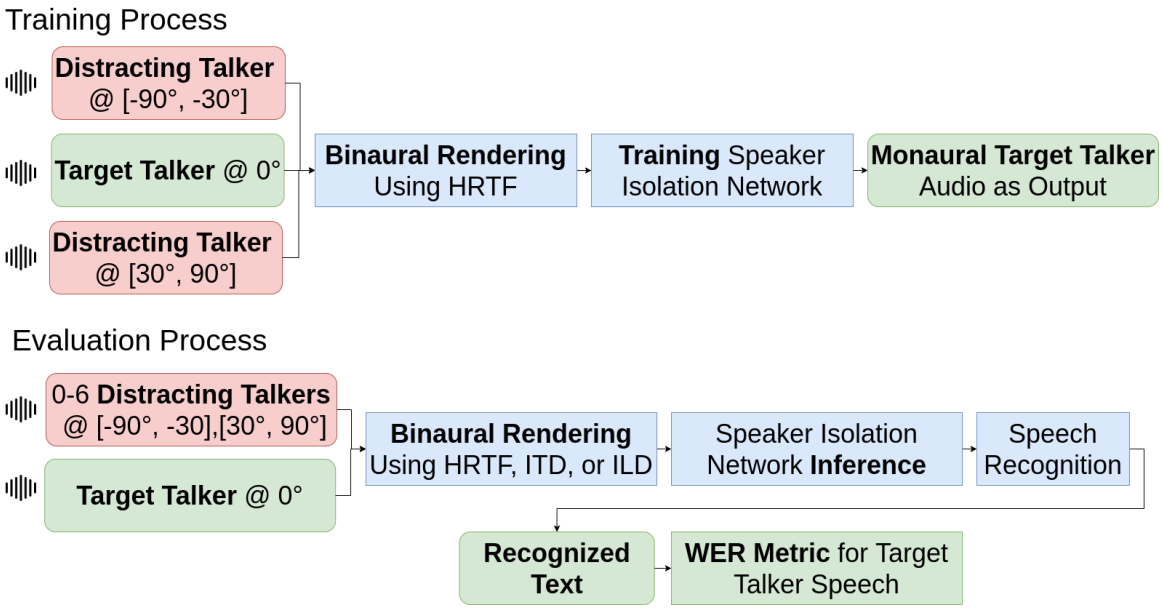

Figure 2.1: **BSINet training process.** The training process starts with rendering the target talker at 0°, and distracting talkers rendered randomly to the left at -90°, -60°, or -30°and to the right at 30°, 60°, or 90°. The speaker isolation network was then trained to isolate a monaural channel containing the target talker's speech. For evaluation, the target and distracting talkers were spatially rendered, then inference was performed using the speaker isolation network, and the resulting isolated audio was used as input for a speech recognition system.

Cocktail Party Dataset Generation

To create auditory scenes with multiple talkers rendered at different azimuthal angles away from the midline, we rendered speech from the CSTR VCTK Corpus [63] with a binaural head-related impulse response (HRIR) dataset that has HRIRs recorded at different spatial angles. We used the SADIE 2 HRIR Database [2]. To render a talker to a particular angle, we convolved the monaural speech of that talker with the left and right HRIR corresponding to the specific desired azimuthal angle. Each talker in the scene was convolved with the HRIR corresponding to its assigned spatial angle and the different left and right channels were summed together to create the final multi-talker auditory scene. We always rendered the target talker at 0°and distracting talkers randomly at -90°, -60°, -30°, 30°, 60°, 90°. The training dataset contained 2 distracting talkers, but we tested from 0 to 6 distracting talkers during evaluation.

Evaluation Dataset to Isolate Contribution of ITD and ILD

The human auditory system uses two main cues for spatial hearing, the Interaural Time Delay (ITD) and Interaural Level Difference (ITD). To evaluate how these cues might be learned by our network, we created distinct evaluation datasets in which sentences were rendered using either the previously described HRIR method, or only an ITD rendering method, or only an ILD rendering method. The ITD rendering method consisted of copying the monaural speech signal into left and right channels, and delaying the left or right channel using the appropriate delay for the corresponding angle. The delay Δt was calculated using the following equation

$$\Delta t = \frac{r * (sin(\Theta) + \Theta)}{c} \tag{2.1}$$

where r is the head radius, Θ is the arrival angle of the sound in radians, and c is the speed of sound.

For the ILD rendering method we estimated the ILD (in dB) from the HRIR dataset

using the method described in [61]. This consisted of calculating the ILD for each HRIR using the following formula

$$ILD = 20 * log_{10} \frac{|H_l(f)|}{|H_r(f)|} \tag{2.2}$$

where f are the 30 Equivalent Rectangular Band spaced frequencies from 20 Hz to 20000 Hz, and $H_l(f)$ and $H_r(f)$ represent the amplitude of the HRIR at a given frequency for the left and right ears respectively.

2.2.2 System Architecture

Our system is a 1D dilated convolutional neural network that was trained on raw audio, and was based on FFTNet [26]. The original FFTNet architecture described by [26] is depicted in Figure 2.2, and consists of multiple layers that clip their input into two halves, apply a transformation to each half, and then sum and transform the two halves. Specifically, when given a 1D input defined as $x_{0:N}$, an FFTNet layer splits the input into a left half $x_l = x_{0:N/2}$ and a right half $x_r = x_{N/2:N}$ and applies separate convolutions to them:

$$z = W_l * x_l + W_r * x_r \tag{2.3}$$

where W_l and W_r are the weights for the 1x1 convolutions for the left and right side respectively. Once the left and right sides are combined, a ReLU transformation is applied followed by another 1D convolution, and a final ReLU as shown in the following equation $z' = ReLU(conv1x1(ReLU(z)))$ where conv1x1 is the 1D convolution operation.

By stacking multiple FFTNet layers, the full network takes multiple audio samples as input and outputs a single sample. The network from the original FFTNet paper predicted a current audio sample when given previous samples, and could be used as a vocoder, when given previously generated samples and an auxiliary condition such as the Mel-frequency cepstral coefficients (MFCCs).

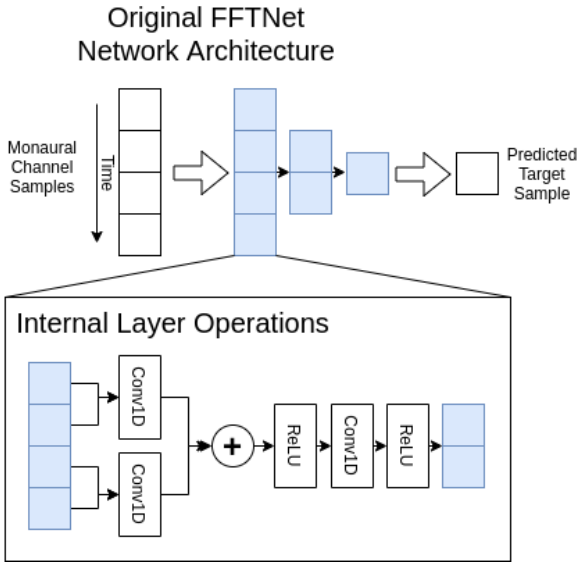

Figure 2.2: **Original FFTNet architecture** due to [26]. The operations inside each layer are as follows: a given input is first divided into two halves; each half is then passed through a different 1x1 convolution layer and then summed together. This summed output is passed through a ReLU function, another 1x1 convolution, and a final ReLU.

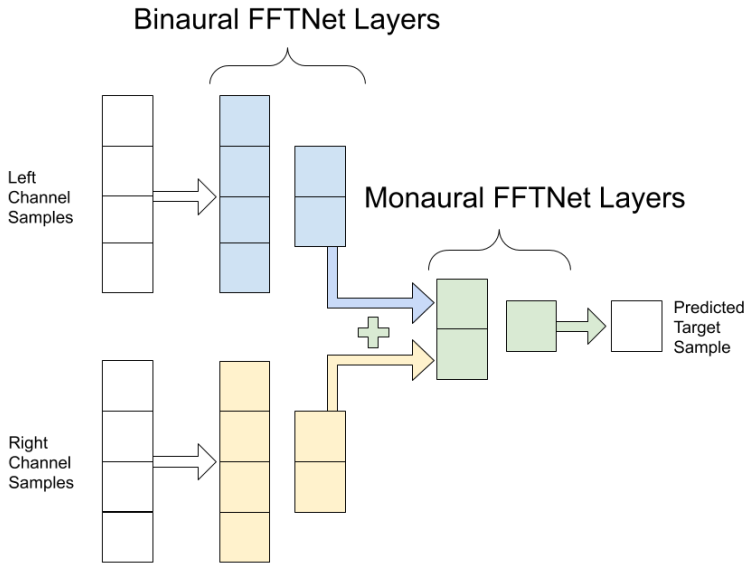

Figure 2.3: **Binaural Speaker Isolation Network (BSINet).** The network consists of two separate layers for the left and right channels, the output of the left and right layers are then combined together and fed into a third set of layers. The network outputs monaural samples of audio for the target talker.

One major difference between our architecture and the original FFTNet architecture is that our network takes binaural input from two ears. This novel architecture is inspired by the early auditory pathway found in mammals, and is shown in Figure 2.3. The model consists of initial binaural layers that process the left and right channels independently. These independent left and right layers are analogous to the Bushy Cells in the mammalian auditory system, which span inputs from the basilar membrane via the auditory nerve [15]. The output of these left and right networks are then combined into a third monaural network that emulates the binaural integration function of the Superior Olivary Complex. This merging of the left and right binaural layers is implemented by combining their outputs during the summation step in the first monaural FFTNet layer. Given the output of the left layers are defined as x_a and the output of the right layers are x_b, then the transformation of the first monaural layer before the ReLU step is defined by the equation:

$$z = W_{al} * x_{al} + W_{ar} * x_{ar} + W_{bl} * x_{bl} + W_{br} * x_{br} \tag{2.4}$$

where x_{al} and x_{ar} are the left and right halves of x_a, x_{bl} and x_{br} are the left and right halves of x_b, and W_{al}, W_{ar}, W_{bl}, W_{br} are the corresponding conv1x1 weights. The last two layers of the network after the monaural FFTNet layers are a fully connected layer and a softmax layer containing 256 units.

The final softmax layer is used because the network takes audio samples that are converted from floating point numbers to 8-bit integers using the following μ-law encoding function f, where $\mu = 255$, x is the floating point audio sample in the range $-1 <= x <= 1$, and $sgn(x)$ is the sign function.

$$f(x) = sgn(x)\frac{ln(1+\mu|x|)}{ln(1+\mu)} \tag{2.5}$$

This converts the network to a classification network with 256 classes as output, corresponding to the 256 possible values for an 8-bit integer.

Another difference between our architecture and the original FFTNet architecture is that the original FFTNet was a vocoder system and as such could only produce one sample at a time. The produced samples would then be fed back iteratively into the network to generate the next consecutive sample. In our architecture we already have the entire input sequence fully defined and this enables us to generate multiple output samples in one shot, similar to the adaptation of WaveNet for speech denoising shown in [44].

2.2.3 Training Process

Speaker Isolation Task

The network takes raw audio samples containing multiple talkers and isolates the target talker's speech, outputting a single channel of audio corresponding to the target talker. i.e. when trained on input samples $x_{1..n}$ the network learns to output samples $y_{1..n}$, where x is two-channel audio containing multiple talkers and y is the audio of the target talker. The input, output, and target audio samples are encoded using the previously described μ-law encoding. This converts the speaker isolation task into a classification problem. The output for each sample is a one-hot encoded vector of size 256 and the network is trained using the Negative Log Likelihood Loss function.

We trained BSINet on complex auditory scenes consisting of sentences from the CSTR VCTK Corpus[63], a dataset of speech from 109 native speakers of English. The sentences were spatially rendered using the previously described preprocessing steps.

Evaluation System

For evaluation of BSINet, we use the isolated audio from the network as input for a speech recognition system. In our experiments we used Mozilla's open-source DeepSpeech implementation[36][19]. We used version 0.6.1 of the included pre-trained acoustic and language models. Performance was then evaluated using the Word Error Rate (WER) between the output of the speech recognition and the ground-truth transcript of the target talker.

We also evaluated the isolated audio from the networks using an evaluation metric called the Source-to-Distortion Ratio (SDR) as defined in [57]. The Source-to-Distortion Ratio is a commonly used metric for evaluation of speaker separation systems, and is defined by the following equation

$$SDR = 10 * log_{10} \frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{2.6}$$

where s_{target} is a version of the target talker's speech and e_{interf}, e_{noise}, and e_{artif} are estimated error terms corresponding to interfering sources, noise, and artifacts respectively. In order to use SDR to evaluate the performance of the network, the SDR is calculated for the original spatially rendered audio as well as the isolated audio. Then the BSINet's performance can be quantified by calculating the Δ SDR improvement between the original and isolated audio.

2.3 Results

The following section first outlines the experimental setup of the proposed BSINet as well as a monaural baseline network. Then, results are shown for each network using the Word Error Rate (WER) metric for evaluation.

2.3.1 Network Experimental Setup

The BSINet we trained had two initial binaural layers followed by nine monaural layers, giving the network a temporal receptive field of 2048 audio samples. Each layer had 128 units. The network was trained on sequences of 5000 samples x 2 channels corresponding to the left and right ear. Given this input, the network was trained to isolate a single channel of audio containing the target talker's speech. We used a batch size of 5 sequences and the Adam[29] optimizer with a learning rate of 0.001, similar to the original FFTNet paper. As a baseline we also trained a standard FFTNet with 11 layers on single-channel audio. We trained each network for 120 epochs to ensure convergence.

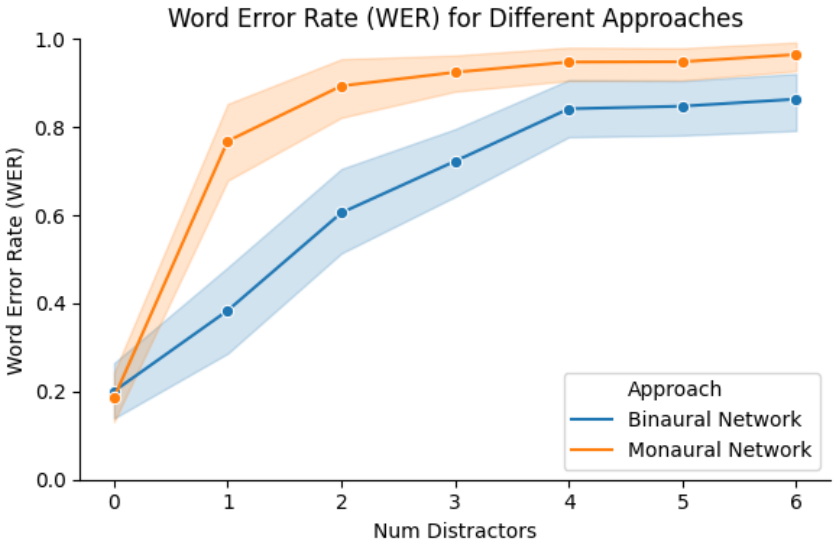

Figure 2.4: **Word Error Rate (WER) performance for different networks.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of approach with F = **83.55** and a main effect of number of distractors with F = **117.18**.

2.3.2 WER Evaluation

The first test we performed on the different networks was an evaluation using the Word Error Rate (WER) metric. The test audio was run though both the BSINet as well as the baseline network. The output audio from each network was then converted to text using the Mozilla DeepSpeech system [36]. This text was finally compared to the ground-truth transcripts using the previously described WER metric. The results are shown in figure 2.4.

2.3.3 SDR Evaluation

The next evaluation we performed was to test the SDR improvement on both of the different networks. To do this we used the python mir_eval toolbox [41]. The results are shown in table 2.1. The Binaural Network shows excellent SDR improvement across all levels of increasing distractor set size, whereas the monaural network only shows improvement when there are no distractors present. This indicates that the monaural network did not learn to isolate the target talker although it was still able to remove the effects of the HRTF when no distracting talkers were present.

Table 2.1: **SDR Improvement for Monaural and Binaural Network**

Δ SDR (dB)							
Network	# of Distractors						
	0	1	2	3	4	5	6
Binaural Network	+15.94	+13.21	+10.91	+7.67	+5.93	+4.92	+4.79
Monaural Network	+13.07	-0.16	-1.32	-1.28	-1.44	-2.17	-2.01

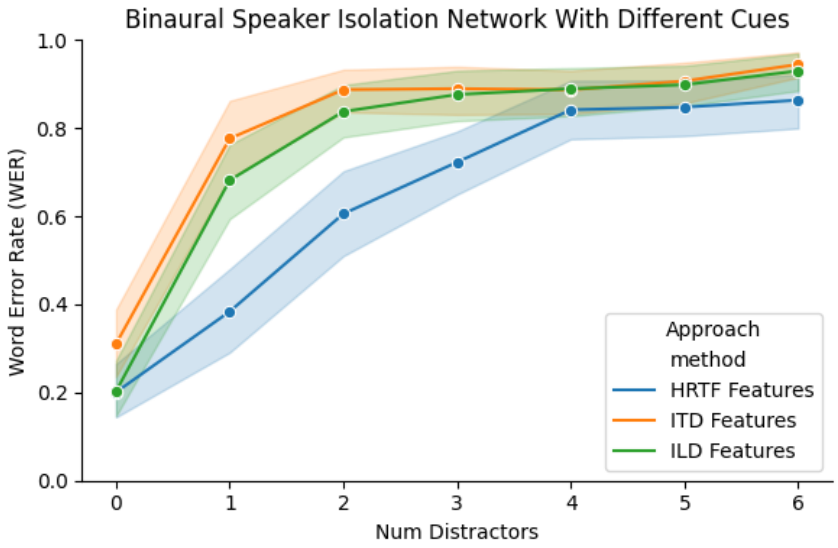

Figure 2.5: **WER performance for BSINet with different spatial cues.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of feature type with F = **42.89** and a main effect of number of distractors with F = **148.32**.

2.3.4 WER Evaluation on Different Cues

The next step in our evaluation was testing the WER performance of the network when using different cues for rendering the talkers. Figure 2.5 shows the WER results for the network. The network performed best when it had access to the full HRTF features relative to ILD features or ITD features only.

2.4 Summary

This chapter has demonstrated a neural network architecture called BSINet; a network inspired by the early auditory pathway in the mammalian brain. As outlined in figure 2.4 the

network based on binaural cues dramatically outperformed a monaural network, and was robust to an increasing set size of distractors, unlike the monaural network. An experiment that evaluated the extent to which BSINet relies on the different binaural cues, such as ITD and ILD, was also presented. The implications and applications of the success of BSINet are elaborated in Chapter 4 below.

The robustness to distraction of the BSINet suggests human-like ability to focus attention on a single talker. To further explore this, the next chapter of this thesis tests human participants on the same dataset used in to evaluate BSINet. We tested the prediction that BSINet, by incorporating binaural spatial cues, has learned the human-like ability to focus attentional selection on a target talker voice.

There is a subtle point to note in the interpretation of the second experiment's results: rendering a talker at an angle using only an ILD or ITD cue means that there is conflicting spatial evidence from the other cue that the talker is at 0°. For example, if a talker is rendered using only ILD, this means that the time difference between the two channels is zero, which corresponds to a spatial location of 0°. Vice versa, a talker rendered using only ITD has a balanced level difference between the two channels, which also implies a spatial location of 0°. This is an important subtlety, as a mismatch between ITD and ILD most likely never occurs in real-world free-field listening. For this reason it is unclear from the present results whether the lack of independence of ITD and ILD cues is critical, or indeed whether it pertains in the case of human listeners. This point is further explored in Chapter 3.

Chapter 3

Comparison of Network and Human Performance Across Cues

3.1 Introduction

Chapter 1 of this thesis has outlined the major cues the mammalian brain uses to isolate speech in multi-talker environments, and chapter 2 proposed a neural network architecture called BSINet based on these cues and evaluated the extent to which BSINet relies on each cue.

The fact that BSINet is based on these cues raises a few questions; namely, how similar is the performance of the network to human performance and does the network rely on binaural cues in the same way humans do? This chapter contains an experiment that tests human participants on the same multi-talker dataset from chapter 2. The results of this experiment allow a direct comparison between the human participant performance and BSINet, as well as evaluating the extent to which the human participants rely on each spatial cue.

This chapter starts by delving into the similarities and differences between neural networks and the mammalian brain and what can and cannot be inferred through comparisons between the two. Then, an experiment using the same test data from chapter 2 on human participants is presented. Finally, the network's performance is compared to that of human participants who performed the same speech recognition task under the same manipulation of distractor set size.

3.1.1 Comparing Neural Networks and the Mammalian Brain

Neural networks have shown impressive performance when applied to problems in many different domains, such as image recognition [42], language modelling [5], speech recognition [19], and many others. The earliest developments in artificial neural networks, called perceptrons, were directly inspired from observations of learning in the visual system [48][35]. This is of note, especially since other developments in neuroscience have led to corresponding improvements to artificial neural networks. For example, the mechanism of visual attention in the human brain has led to improvements in neural networks that perform image captioning as shown in [62]. The attention mechanism has also vastly improved networks that solve various problems in the natural language processing domain such as translation, summarization, question-answering, as shown in [56][40].

Although insights from the brain have enabled improvements in neural network performance, an important aspect to note is that this does not imply that neural networks are learning to perform tasks in the same way the human brain does. An example that illustrates this is a problem known as the credit assignment problem. As outlined in [45], the credit assignment problem describes the difficulty in assigning weight to different connections in systems that are composed of hierarchical networks containing multiple layers. In order for such a system to learn, a weight must be assigned to each connection in the network based on how much it contributed to the correct output. The approach used to solve this problem in artificial neural networks is known as backpropagation [49]. While there are theories that attempt to explain how a process similar to backpropagation might happen in the brain [50], having a separate backward pass of updating weights based on error is still biologically implausible as outlined in [46][16]. This is an example of how a neural network and the human brain might arrive at the same computational results by taking different computational approaches.

Even though the process of backpropagation itself is biologically implausible, in some cases the networks trained through backpropagation have been shown to replicate processes

that happen in the brain. A few examples of this can be found in research such as [3], in which a recurrent neural network learned representations that were similar to the grid cells found in the hippocampus, and [39] in which internal neurons from a deep neural network trained on an image classification task showed similar translation-invariant selectivity to boundaries as neurons found in visual processing area V4 in primates. An area of research that takes such comparisons even further is Brain-Score [51] which quantifies the amount of similarity between artificial neural networks and the brain on visual object recognition tasks. The Brain-Score paper defined a benchmark based on both neural recordings from primates and behavioural object recognition tasks. This enables researchers to run their image classification networks on the same benchmark to test the biological plausibility of their models.

Another fascinating example is found in a neural network architecture called PredNet [33]. PredNet is an architecture based on predictive coding, which is a model of how the brain processes information internally. The authors in [60] trained a PredNet model to predict future frames of video, and when the network was presented with a visual illusion it produced the same illusory predictions that the human visual system makes. This makes PredNet an interesting analogue in the visual domain to the research in this thesis as PredNet is inspired by the architecture in the brain, similar to the BSINet architecture from chapter two. We take the same approach that [60] took when they compared their network architecture to the visual system in the brain by using specially crafted illusory inputs. The research in the following chapter will compare BSINet to behavioural data from human participants in order to test how similarly BSINet performs to human participants when given various binaural cues. This provides a simple methodology for us to benchmark how similar our network is to humans on various cues. We also provide the dataset in Appendix A. This will enable other researchers to also test their networks against the human participant data.

3.2 Methodology

Our goal was to evaluate the performance of our network against human participant performance in a complex ”cocktail party” auditory scene. To do this we recruited human participants to perform the same speaker isolation and speech recognition task. Eight undergraduate students were recruited from the University of Lethbridge and received course credit for their participation. All participants were neurologically normal and reported normal hearing in a pre-experiment questionnaire. The participants were able to adjust the playback volume to comfortable levels at the beginning of the experiment. The experiments and procedures adhered to the Declaration of Helsinki and were approved by the University of Lethbridge Institutional Review Board.

3.2.1 Human Participant Experimental Setup

The experiment was conducted in an acoustically damped room. The participants were seated in a standard rolling office chair at a desk with a computer for stimulus presentation. Participants were wearing Sennheiser over-ear headphones and listened to audio from the previously described evaluation dataset. Stimulus presentation and response recording were both performed using the PsychoPy package[38] for Python. The participants were instructed to type a sentence corresponding to the speech they heard from the talker perceived to be in the center at 0°. After being given these instructions, they were played an initial example file in order to set the volume to a comfortable level. The example file contained multiple distracting talkers and the participants were not instructed to type a response for this example file.

As in our evaluation of our network in chapter 2, each trial consisted of a presentation of target speech plus distraction ranging from zero to six distracting talkers. We had three conditions for the rendered audio: the ITD condition provided only binaural time lag cues, the ILD condition provided level-difference cues, and the Head Related Transfer Function (HRTF) condition provided the entire complement of binaural cues available in the Sadie 2

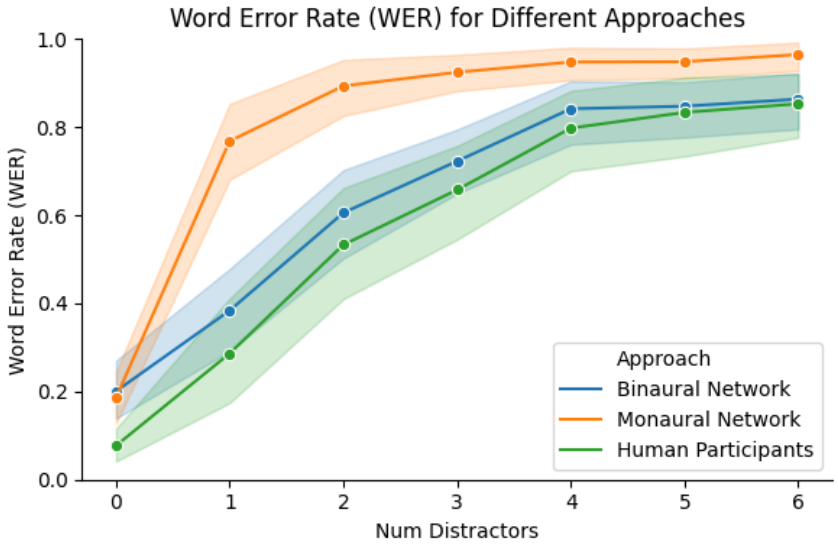

Figure 3.1: **Performance comparison between human participants and BSINet.** The coloured bands show the 95% confidence interval of the results. A two-way ANOVA showed a main effect of approach with F = **61.44** and a main effect of number of distractors with F = **141.60**.

HRTF models. We used 5 sentences for each distractor # x rendering condition for a total of 105 sentences per participant. The target sentence in each trial was unique across all trials, and all sentences during each trial came from different VCTK talkers.

3.3 Results

The following section contains the results from the human participants. The results are shown using the Word Error Rate (WER) metric, with the data from the neural network experiment in chapter 2 plotted alongside for comparison.

3.3.1 WER Evaluation on HRIR Features

The first test we performed was the evaluation of each network and the human participants on their Word-Error-Rate (WER) performance on the HRIR rendered subset of audio, shown in Figure 3.1. The BSINet performed similarly to human participants when one or more distractors were present. By contrast, the monaural network exhibited a much steeper rise in WER as distractor set size increased.

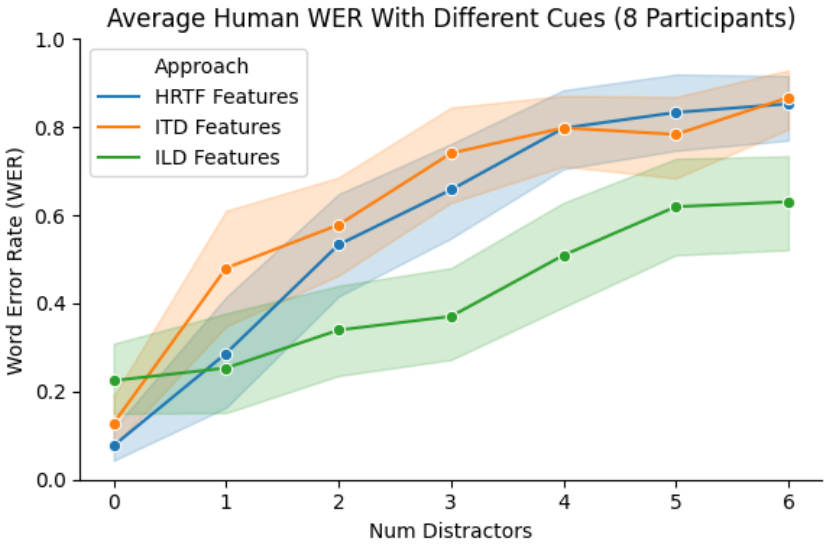

Figure 3.2: **Human participant performance with different spatial cues.** The coloured bands show the 95% confidence interval of the results across all participants. A two-way ANOVA showed a main effect of feature type with F = **30.85** and a main effect of number of distractors with F = **63.85**.

3.3.2 WER Evaluation on Different Cues

The next step in our evaluation was testing the WER performance of the participants and the network when using different cues for rendering the talkers. Figures 3.2 and 2.5 show the WER results for the participants and network respectively. The coloured bands show the 95% confidence interval of the results across all participants. The participants performed best on the ILD rendered subset when one or more distractors were present. This was in contrast with the BSINet, which performed best when the full HRTF cues were available, and similarly with either ITD or ILD cues.

3.4 Summary

This chapter started by exploring existing comparisons between neural networks and the human brain. These comparisons are valuable because the first neural networks were initially inspired by learning in the visual system (e.g. see [48] and [35]), and insights from neuroscience have led to improvements in neural network performance on a variety of tasks.

Next, an experiment that compared the performance of BSINet to human participants was presented. The results of this experiment are interesting as the human participant's performance is comparable with BSINet's performance on the HRTF rendered audio. Historically, human perceptually performance has been regarded as a "gold standard" to which machine vision and hearing systems were compared. In this sense, BSINet performs exceptionally well, rivaling the performance of human listeners. More recently, artificially intelligent systems have begun to exhibit "superhuman" abilities, albeit within highly constrained domains (e.g. Google's AlphaGo and IBM's DeepBlue). Thus it is an interesting unanswered questions as to whether human performance should be regarded as the optimal benchmark for artificial perception systems such as BSINet.

The experiment in this chapter also shows some differences between the network and participants, specifically in Figure 3.2, which looks at human participant performance on the different cues of ITD and ILD. This figure shows that humans showed better performance on the ILD cue subset of the data than the HRTF and ILD subsets, which were similar to each other. This would imply that some aspect of the HRTF and ITD rendering process potentially detracts from the human participant's ability to isolate the target talker. Another explanation would be that humans do in fact rely more heavily on the ILD cue than other cues. Chapter 4, Section 4.1.1 further explores these potential explanations.

Chapter 4

Discussion and Future Directions

4.1 Discussion

The mechanisms by which the human brain unmixes the complex auditory scene and selectively listens to a single voice have been of interest to psychologists and cognitive neuroscientists for almost a century. Similarly, the applied problem of computationally unmixing and understanding target speech in noisy environments represents a challenging obstacle in the development of auditory artificial intelligence for human-robot interaction and control of autonomous systems.

In this research we have demonstrated BSINet, a network that improves speech recognition performance in complex auditory scenes. The network learned to use binaural cues to resolve a target talker on the azimuthal midline from a mixture of distracting talkers at various arrival angles. We compared the performance of this binaural network to a comparable monaural network with respect to its robustness in the face of an increasing set size of distracting talkers. This manipulation of set size was inspired by related foundational work in the psychology of visual selective attention e.g. [53], which characterized how it becomes increasingly difficult to find and resolve certain unique targets (known as "non-popout" targets) amid visual clutter. Although systematically measuring perceptual performance as a function of set size is quite common in the visual search attention literature, it is rarely used to gauge performance on auditory tasks with realistically complex auditory scenes (see e.g. [18]).

It is well known that human listeners use space as a dimension to separate sound sources

in complex scenes - a phenomenon known as spatial release from masking [22][64][20][28][32][1]. Although the ITD and ILD binaural cues are known to provide the basis for sound source localization, it is unclear how either or both of these cues might contribute to the spatial unmixing of speech performed by the brain in the case of multiple distractors, or indeed how these cues might be utilized to improve artificial systems that must recognize speech in noisy environments. By varying the set of distractors, we showed in figure 3.1 that a traditional monaural network is not robust to increasing distractor set size. Word error rate increased dramatically with inclusion of only a few distracting talkers. By contrast, our binaural network exhibited a degree of robustness to increasing distractor set size. In fact, our network performed remarkably similar to human listeners performing the same task.

Figure 3.1 shows that a binaural network outperforms a monaural network as distraction increases, but does not indicate which binaural cues the network has learned to use to resolve the auditory scene. To investigate this, we rendered distractors either exclusively with ITD or ILD cues, or with the entire head-related transfer function, which provides both ITD and ILD cues along with other impulse-response modulations associated with propagation of sounds in free field around the head. In interpreting the result, it is important to recognize a subtlety of this approach: It is possible to impose ITD or ILD cues independently so as to render a sound to an arrival angle off of the midline. This is done, respectively, by systematically varying the interaural time lag (ITD) or modulating the interaural sound intensity difference (ILD) for the left and right channels for each sound in the scene. However, doing so necessarily leaves the other cue unaffected. Thus in this sense, a distractor may be rendered to an azimuthal arrival angle other than zero degrees in ”ITD space”, but it is left at zero degrees in ”ILD space”. Vice versa, a distractor may be rendered off the midline by ILD, but left on the midline with respect to ITD. The question that can be asked is thus rather nuanced: has the network (or indeed a human listener) learned to represent ILD and ITD as independent dimensions that can be used sufficiently to unmix sound sources?

In figure 2.5 we show that BSINet exhibits the most robustness to distraction when given

the entire complement of binaural cues available in the HRTF. When the distractors were rendered spatially distinct from targets by only ITD or ILD, the network was less robust to distraction, with performance more like that of the monaural network shown in figure 3.1. This is perhaps unsurprising since the binaural network was trained on a dataset rendered with precisely the same HRTF. However, it suggests that the computations performed by BSINet do not generate independent representations of ITD and ILD ”spaces”. Interestingly, human listeners showed a different pattern of results. Humans exhibited robustness to distraction when distractors were rendered spatially with ILD alone, but exhibited somewhat less robustness when only the ITD cue or the HRTF was provided. This suggests that the human auditory system makes use of auditory space in rather different ways that were not exploited by our network. Likewise, the network appears to exploit aspects of the HRTF that are not provided by rendering the distractors solely with ITD or ILD. Understanding this discrepancy promises to reveal novel approaches for speech denoising with more human-like characteristics.

4.1.1 Comparison with Biological Systems

The similarity in performance between BSINet and human participants is interesting as it shows that the network makes use of binaural cues to isolate a single talker effectively from a complex scene. The differences between the network and human participants become more apparent when looking at the performance on specific cues such as ITD and ILD. Figure 3.2 shows that the human participants showed better performance using the ILD cue alone than either the ITD or HRTF cues. In fact, performance with distractors rendered with the complete HRTF and exclusively ITD was qualitatively similar, and significantly worse than performance when distractors were rendered exclusively with ILD.

One explanation for this may be purely methodological: small differences in time delays between the ears can correspond to large differences in the apparent sound arrival angle, especially around 90 deg from the midline. The individual participants in our study would

have had slightly different distances between the ears than the Sadie II dataset provides, giving rise to slightly aberrant binaural delays for each individual participant. In fact in the case that a participant's head was actually narrower than the Sadie II head, a sound rendered at 90 deg from the midline would feature binaural ITD slightly beyond what would be physiologically possible. This means unfamiliar ITD cues by themselves, and also the ITD component of the HRTF cues, might actually detract from the human participant's ability to isolate the target talker in the ITD-only and HRTF conditions. The network on the other hand was trained using the same HRTF and ITD used during testing, which possibly enabled it to make better use of the ITD cue more effectively.

An alternative explanation for the pattern of results in Figure 3.2 is that human listeners do actually represent ITD and ILD "spaces" independently and most effectively use the ILD cue to solve the cocktail party problem. A few previous studies have considered the independence of ITD and ILD in spatial hearing for the purpose of spatial release from masking by competing sounds. Culling 2004 [11] modified a prerecorded head-related transfer function (HRTF) to contain only ITD or only ILD cues. They used this cue to render distracting sounds to peripheral locations while target speech was located at zero degrees (i.e. on the azimuthal midline, as in our study). They concluded that in general, both cues were independently sufficient to mediate release from masking when all the distractors were located in the same hemifield. However, when distractors were present in both hemifields, relatively little release from masking was obtained for the ILD-only condition, while ITD did allow release from masking. Kidd et al. [28] approached the question differently by leveraging the fact that ITD and ILD are believed to operate over different regions of the frequency domain. As described by the duplex theory, the auditory system is thought to make best use of ITD for low frequency sounds (i.e. below about 2000Hz) and ILD for high-frequency sounds (i.e. above about 1000 Hz). By bandpass filtering the distracting stimuli, they found that spatial release from masking was still evident over all frequency ranges, although somewhat more effective for low-frequency filtered distractors, suggesting

like Culling (2004) that the ITD cue is more useful in spatial release from masking. More recently, Glyde et al. [13] also used HRTF modified to contain only ITD or only ILD cues. They found that both ILD-only and ITD-only conditions enabled substantial release from masking, however the ILD-only condition allowed better performance than the ITD-only condition - suggesting in contrast to Culling (2004) and Kidd (2010) that the brain more effectively makes use of ILD. Our results agree more closely with those of Glyde et al. (2013) in that the ILD-only condition allowed for the most robustness from increasing set size of distractors among human listeners. However, our network seems not to make preferential use of one or the other cue.

4.1.2 Comparison with Previous Computational Approaches

Most existing systems aimed at solving the cocktail party problem are either multi-channel systems based on traditional techniques such as beamforming and Independent Component Analysis or are single-channel systems that make use of recent advances in deep learning. In this research we have demonstrated a Binaural Speaker Isolation Network that improves speech recognition performance in complex auditory scenes. This system is important because of its excellent performance when several distracting talkers are present (without substantially sacrificing performance on ”clean” speech). This system comes with other benefits as well as some drawbacks when compared to traditional approaches. The main benefit of our system compared to other multi-channel systems is that it only requires the use of two microphones. Both beamforming and ICA work better as the number of microphones increase, and ICA specifically requires $n+1$ microphones in order to separate n sources. Another benefit to our system is that it can learn to exploit the full complement of features available in the HRTF. For example, if the BSINet is deployed on a robot with physical components that block or shape sound propagation in free field, then the robot's HRTF is probably a non-trivial one with many useful features the system can use. One drawback to the BSINet architecture however is that in order to change the microphone

configuration, the network must be retrained from scratch on the new configuration. In contrast, both beamforming and ICA-based systems can be configured for any microphone setup during runtime.

4.2 Future Work and Applications

This thesis has demonstrated the basic neural network architecture for BSINet. There are a variety of improvements that could be made to the neural network architecture, future experiments that can further explore the similarities and differences between BSINet and the human brain, as well as various applications of the network. This section starts by outlining future improvements to the network and then delves into potential experiments and applications of the network to real-world problems.

4.2.1 Architecture Improvements

There are improvements to BSINet that we hope to make in the future. A first potential improvement to the network architecture could be a thorough exploration of possible network parameters, such as testing various numbers of units in each layer, as well as testing networks with varying numbers of binaural layers before the merging step. In this research, however, our experiments were performed on a machine with an Nvidia 1070 TI graphics card, and training the BSINet took several days. This means a rigorous parameter search of different network configurations would probably involve running many experiments in parallel on different GPUs.

BSINet was trained to isolate a target voice on the auditory midline from distractors located at various azimuthal angles. Listening to a talker directly in front is a common situation for human listeners, but we can also direct our auditory attention toward sources away from the midline. Thus, another useful improvement would be to train a BSINet that is globally conditioned with a target spatial angle so that the network's attention can be steered, which would enable the network to isolate the speech of a talker at any angle. This

would be useful for applications such as humanoid robotics, in which the robot could attend to a target talker in its environment while orienting its head in a different direction. However, this would require somewhat sophisticated steering of the beam of attentional focus to solve the kinematic transforms needed to account for the pose of the robot. Finally, a further performance improvement to be gained in overall WER performance is to customize the speech recognition system model. Specifically, the DeepSpeech model used in this research was trained on clean audio and training or even just fine-tuning it on isolated speech from the output of the BSINet could potentially lead to overall performance gains.

4.2.2 Future Experiments

There are several experiments that might lead to a better understanding of how the BSINet achieves its robustness to distraction, mostly through a more in-depth comparison of the auditory processing in the BSINet and the human brain. One type of experiment could be an exploration of the neural activations in the hidden layers of the network. A specific example of this would be an experiment that attempts to reverse-engineer the early feature representations the network has learned. This would be valuable since the network is given raw audio instead of commonly used features such as MFCCs or spectrograms. Such an experiment could shed light on whether or not the initial layers of the network perform a similar frequency decomposition to the one that happens in the basilar membrane in mammals.

Another future experiment could test whether or not there are neurons in the trained BSINet that are tuned to specific ITDs and ILDs during training, similar to neurons in the SOC in the mammalian auditory system. Such neurons would be an interesting discovery, and could potentially provide insight into the specific computations BSINet is using to isolate the target talker's speech. This would allow a more direct comparison with the neural architecture of the brain. An extension of this could be the development of a metric similar to the brain-score metric defined in [51]. Although collecting neural recordings from the

SOC in primates is challenging, a similar comparison using electroencephalography and the neural predictivity metric also outlined in [51] could lead to a more robust comparison between humans and neural networks on speaker isolation tasks.

4.2.3 Applications

The research in this thesis can be leveraged to solve a multitude of real-world applications. BSINet currently isolates the speech of talkers at 0°; if BSINet was processing speech from a physical device this would correspond to isolating the talker directly in front of the device. As outlined previously in the future experiments section, a future version of the network would also be able to direct its attention towards a given target angle. These features mean that the network could be used to improve physical devices such as hearing aids or auditory augmented reality, giving the end-user the ability to spatially isolate the speech of a person they are having a conversation with, potentially even in complex auditory environments such as restaurants and parties. Another similar application of BSINet would be to improve the performance of voice assistants such as Google Home, Alexa, and others. This application would also require other technologies such as speaker direction-of-arrival calculation, since the device would need to steer the BSINet in an automated fashion. Another similar use case would be enabling humanoid robots to selectively direct their auditory attention towards target talkers. Such robots are an interesting use case as one possibility would be to use facial recognition and image processing to determine the target angle of a desired target talker in order to perform the speaker isolation.

4.3 Conclusion

Humans have the remarkable ability to understand speech in noisy environments with many distracting talkers, a problem that has been a difficult one to solve computationally. This thesis demonstrated a neural network architecture called BSINet, that isolated the speech of a single talker from input containing a target talker and multiple distractors.

BSINet was inspired by the early auditory pathways found in the mammalian brain. The network was evaluated using both the WER metric and the Δ SDR improvement across a dataset containing a varying number of distractors. BSINet's performance was found to be excellent, even when multiple distracting talkers were present in the auditory scene. This research also evaluated network performance using a dataset containing different spatial rendering cues such as the Interaural Time Delay (ITD) and Interaural Level Difference (ILD) and Head Related Transfer Function (HRTF) rendering. This was used to compare human participant performance to BSINet's performance, and was also released online. This will enable future researchers to compare the results of their networks to the human participant data.

The comparison between the network and human participants highlighted the similarity in performance between the two on HRTF rendered audio. It also outlined some underlying differences, such as how the human brain could independently make use of ILD cues to isolate the speech of a target talker, whereas the BSINet has not learned to do so. In fact, human participants performed better on the ILD only condition. This raised an important question of whether this difference is actually due to humans being able to more effectively use the ILD cue to solve the cocktail party problem, or whether the small variations in interaural distance led to decreased performance on all audio directly or indirectly using ITD cues as part of the rendering. Future experiments could be designed in order to explore this question more in-depth. Overall, the research in this thesis provides significant advancement for solving the challenging cocktail party problem and has the potential to improve technologies such as hearing aids, humanoid robotics, voice assistants among others.

Bibliography

[1] Tanya L Arbogast, Christine R Mason, and Gerald Kidd Jr. The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 117(4):2169–2180, 2005.

[2] Cal Armstrong, Lewis Thresh, Damian Murphy, and Gavin Kearney. A perceptual evaluation of individual and non-individual hrtfs: A case study of the sadie ii database. *Applied Sciences*, 8(11):2029, 2018.

[3] Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

[4] Albert S Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] CE Carr and M Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *The Journal of Neuroscience*, 10(10):3227–3246, 1990.

[7] Zhuo Chen, Jinyu Li, Xiong Xiao, Takuya Yoshioka, Huaming Wang, Zhenghao Wang, and Yifan Gong. Cracking the cocktail party problem by multi-beam deep attractor network. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 437–444. IEEE, 2017.

[8] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, 25(5):975–979, 1953.

[9] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[10] MP Cooke and Guy J Brown. Computational auditory scene analysis: Exploiting principles of perceived continuity. *Speech Communication*, 13(3-4):391–399, 1993.

[11] J. F. Culling, M. L. Hawley, and R. Y. Litovsky. The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *J Acoust Soc Am*, 116(2):1057–65, 2004.

[12] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. Deep recurrent networks for separation and recognition of single-channel speech in nonstationary background audio. In *New Era for Robust Speech Recognition*, pages 165–186. Springer, 2017.

[13] H. Glyde, J. M. Buchholz, H. Dillon, S. Cameron, and L. Hickson. The importance of interaural time differences and level differences in spatial release from masking. *J Acoust Soc Am*, 134(2):EL147–52, 2013.

[14] J. M. Goldberg and P.B. Brown. Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *J Neurophysiol*, 32:613 – 636, 1969.

[15] B. Grothe, M. Pecka, and D. McAlpine. Mechanisms of sound localization in mammals. *Physiological Reviews*, 90:983 – 1012, 2010.

[16] Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *ELife*, 6:e22901, 2017.

[17] D. A. Hambrook, M. Ilievski, M. Mosadeghzad, and M. S. Tata. A bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. *PLoS One*, 12(10), 2017.

[18] D. A. Hambrook and M. S. Tata. The effects of distractor set-size on neural tracking of attended speech. *Brain Lang*, 190:1–9, 2019.

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[20] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005.

[21] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[22] Ira J Hirsh. The relation between localization and intelligibility. *The Journal of the Acoustical Society of America*, 22(2):196–200, 1950.

[23] Guoning Hu and DeLiang Wang. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on neural networks*, 15(5):1135–1150, 2004.

[24] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[25] Lloyd A Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35, 1948.

[26] Zeyu Jin, Adam Finkelstein, Gautham J Mysore, and Jingwan Lu. Fftnet: A real-time speaker-dependent neural vocoder. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2251–2255. IEEE, 2018.

[27] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. *arXiv preprint arXiv:1802.08435*, 2018.

[28] Gerald Kidd Jr, Christine R Mason, Virginia Best, and Nicole Marrone. Stimulus factors influencing spatial release from speech-on-speech masking. *The Journal of the Acoustical Society of America*, 128(4):1965–1978, 2010.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Bryan Kolb, Ian Q Whishaw, and G Campbell Teskey. *An introduction to brain and behavior*. Worth New York, 2001.

[31] J. C. R. Licklider. *Three auditory theories*, volume 1 of *Psychology: A study of a Science*. McGraw, New York, 1959.

[32] JCR Licklider. The influence of interaural phase relations upon the masking of speech by white noise. *The Journal of the Acoustical Society of America*, 20(2):150–159, 1948.

[33] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[34] Yi Luo, Zhuo Chen, and Nima Mesgarani. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4):787–796, 2018.

[35] Marvin Minsky and Seymour A Papert. *Perceptrons: An introduction to computational geometry*. The MIT Press, 2017.

[36] Mozilla. mozilla/deepspeech, 2020.

[37] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[38] Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007.

[39] Dean A Pospisil, Anitha Pasupathy, and Wyeth Bair. 'artiphysiology'reveals v4-like shape tuning in a deep network trained for image classification. *Elife*, 7:e38242, 2018.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[41] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[42] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[43] Lord Rayleigh. Xii. on our perception of sound direction. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 13(74):214–232, 1907.

[44] Dario Rethage, Jordi Pons, and Xavier Serra. A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE, 2018.

[45] Blake A Richards and Timothy P Lillicrap. Dendritic solutions to the credit assignment problem. *Current opinion in neurobiology*, 54:28–36, 2019.

[46] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[47] Nicoleta Roman, DeLiang Wang, and Guy J Brown. Speech segregation based on sound localization. *The Journal of the Acoustical Society of America*, 114(4):2236–2252, 2003.

[48] Frank Rosenblatt. *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[49] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[50] João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in neural information processing systems*, pages 8721–8732, 2018.

[51] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018.

[52] S. A. Shamma, M. Elhilali, and C. Micheyl. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, 34(3):114 – 123, 2011.

[53] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognit Psychol*, 12(1):97–136, 1980.

[54] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, and Chin-Hui Lee. Deep neural network based speech separation for robust speech recognition. In *2014 12th International Conference on Signal Processing (ICSP)*, pages 532–536. IEEE, 2014.

[55] Yanhui Tu, Jun Du, Yong Xu, Lirong Dai, and Chin-Hui Lee. Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 250–254. IEEE, 2014.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[57] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[58] DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pages 181–197. Springer, 2005.

[59] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014.

[60] Eiji Watanabe, Akiyoshi Kitaoka, Kiwako Sakamoto, Masaki Yasugi, and Kenta Tanaka. Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in psychology*, 9:345, 2018.

[61] Kanji Watanabe, Kenji Ozawa, Yukio Iwaya, Yôiti Suzuki, and Kenji Aso. Estimation of interaural level difference based on anthropometry and its effect on sound localization. *The Journal of the Acoustical Society of America*, 122(5):2832–2841, 2007.

[62] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[63] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92), 2019.

[64] William A Yost and George Gourevitch. *Directional hearing*. Springer, 1987.

[65] STEVEN R Young and EDWIN W Rubel. Frequency-specific projections of individual neurons in chick brainstem auditory nuclei. *Journal of Neuroscience*, 3(7):1373–1378, 1983.

Appendix A

Appendix: Cocktail Party Speech Dataset

The cocktail party speech dataset can be found on Github at the following link:
https://github.com/lukeinator42/cocktail_party_audio_dataset

A.1 Dataset Overview

This dataset contains binaurally-rendered audio for testing speech isolation and recognition systems on ”cocktail party” scenarios.

The rendering process is outlined in Chapter 2, Section 2.2.1. Each audio file contains a target talker and 0-6 distracting talkers. The test_set.csv file describes each file, the number of distractors the file contains, the rendering method, and the target transcript.

The human_participant_responses folder contains the participant responses on the test audio. Punctuation has been removed with the exception of the apostrophes/single quote characters, which have been replaced with underscores.

A.2 Data Sources and Licensing

The speech audio in this dataset is taken from the VCTK corpus and rendering using the HRTF cue makes use of the SADIE II Database. The SADIE II Database is made available under the Apache 2 License.

This dataset contains information from VCTK corpus which is made available under the ODC Attribution License. This corpus is also licensed under Open Data Commons Attribution License (ODC-By) v1.0. http://opendatacommons.org/licenses/by/1.0/ http://opendatacommons.org/licenses/by/summary/.