Resynthesizing Volumetric Soundscapes

Low-Rank Subspace Methods for Soundfield Estimation and Reconstruction

by Spencer Russell
M.S., Massachusetts Institute of Technology (2015)
B.S.E., Columbia University (2008)
B.A., Oberlin College (2008)

Submitted to the Program in Media Arts and Sciences, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences at the Massachusetts Institute of Technology.

Submitted May 2020

Author

Spencer Russell
MIT Media Lab
May 1, 2020

Certified by

Joseph A. Paradiso
Alexander W. Dreyfoos Professor
Program in Media Arts and Sciences

Accepted by

Tod Machover
Academic Head
Program in Media Arts and Sciences

Resynthesizing Volumetric Soundscapes

Low-Rank Subspace Methods for Soundfield Estimation and Reconstruction

by Spencer Russell

Submitted to the Program in Media Arts and Sciences on May 1, 2020, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences at the Massachusetts Institute of Technology.

Abstract

Sound and space are fundamentally intertwined, at both a physical and perceptual level. Sound radiates from vibrating materials, filling space and creating a continuous field through which a listener moves. Despite a long history of research in spatial audio, the technology to capture these sounds in space is currently limited. Egocentric (binaural or ambisonic) recording can capture sound from all directions, but only from a limited perspective. Recording individual sources and ambience is labor-intensive, and requires manual intervention and explicit localization.

In this work I propose and implement a new approach, where a distributed collection of microphones captures sound and space together, resynthesizing them for a (now-virtual) listener in a rich volumetric soundscape. This approach offers great flexibility to design new auditory experiences, as well as giving a much more semantically-meaningful description of the space. The research is situated at the Tidmarsh Wildlife Sanctuary, a 600-acre former cranberry farm that underwent the largest-ever freshwater restoration in the northeast. It has been instrumented with a large-scale (300 by 300 m^2) distributed array of 10-18 microphones which has been operating (almost) continuously for several years.

This dissertation details methods for characterizing acoustic propagation in a challenging high-noise environment, and introduces a new method for correcting for clock skew between unsynchronized transmitters and receivers. It also describes a localization method capable of locating sound-producing wildlife within the monitored area, with experiments validating the accuracy to within $5m$.

The scale of the array provides an opportunity to investigate classical array processing techniques in a new context, with nonstationary signals and long interchannel delays. We propose and validate a method for location-informed signal enhancement using a rank-1 spatial covariance matrix approximation, achieving 11dB SDR improvements with no source signal modeling.

These components are brought together in an end-to-end demonstration system that resynthesizes a virtual soundscape from multichannel signals recorded *in situ*, allowing users to explore the space virtually. Positive feedback is reported in a user survey.

Thesis Supervisor:
Joseph A. Paradiso
Alexander W. Dreyfoos (1954) Professor of Media Arts and Sciences

Additional updates and media are available at:
`http://media.mit.edu/~sfr/soundscape_resynth/`

Resynthesizing Volumetric Soundscapes

Low-Rank Subspace Methods for Soundfield Estimation and Reconstruction

by Spencer Russell

Thesis Committee

Certified by_____

Daniel P. W. Ellis
Research Scientist
Google, Inc.

Resynthesizing Volumetric Soundscapes

Low-Rank Subspace Methods for Soundfield Estimation and Reconstruction

by Spencer Russell

Thesis Committee

Certified by_____

Josh McDermott
Associate Professor
MIT Department of Brain and Cognitive Sciences

Contents

Acknowledgments 9

Notation 12

Introduction 14

I Characterizing the Environment

Background: Impulse Response Measurement 21

Audio Infrastructure Overview 25

Speaker Preparation 28

Impulse Response Capture 30

Clock Skew Compensation 31

Results 36

Limitations and Future Work 40

II Acoustic Localization

Background: Cross-Correlation and Delay Estimation 42

Background: TDoA and Nearfield Localization 49

Proposed Spatial Likelihood Function 54

Ground Truth Data Capture 56

Results 57

Limitations and Future Work 59

III Foreground/Background Separation

Background: Subspaces and Matrix Factorization 62

Background: Signal Enhancement and Separation 70

A Low-Rank Filter for Foreground Separation 78

Results 84

Limitations and Future Work 86

IV Rendering the Soundscape

Demo Application 89

Resynthesis Approach 91

Results 93

V Contributions and Conclusion

Contributions 96

Conclusion 99

Glossary 101

Appendix A: User Survey 104

References 106

List of Figures

1 Tidmarsh map with microphone locations 16
2 Equipment Deployed at Tidmarsh 16

3 Time-domain plot of a dirac delta. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal. 21
4 Time-domain plot of an exponential sine sweep. 22
5 Time-domain plot of a maximum-length sequence. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal. 23
6 Time-domain plot of a random-phase multisine stimulus. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal. 23
7 Time-domain plot of a Golay sequence (parts A and B). The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal. 24

8 Microphones deployed *in situ* 26

9 Speaker and microphone used for speaker measurements 28
10 Speaker output SPL against sample gain 29
11 Spectrograms of exponential sine sweep tests 29

12 Speaker set up in the field 30

13 Multiple skewed frames superimposed. Summing them together creates a comb filtering effect 31
14 Comparison of objective functions for period estimation 33
15 Clock Skew Correction Example 34

16 Clock Skew Estimation Histogram 36
17 Clock Skew Estimation Results 36
18 Impulse Response Fit Example 37
19 Impulse Response Per-Band RT60 37
20 Recording Archive Summary 38

21 GCC-PHAT Comparison 48
22 GCC-PHAT Comparison (Zoomed) 48

23 TDOA Hyperbola 50
24 TDOA Hyperbola 50
25 TDOA Hyperbola 51
26 TDOA Hyperbola 51

27 Full Spatial Likelihood Function Output 55

28 Localization Results 58
29 Localization Results (Zoomed) 58

30 Subspace denoising of a sinsoid at a known frequency in white gaussian noise, with 20 random noisy mixtures. (a) shows the noisy mixture and (b) shows the signal estimates (grey) and true signal (black). 62
31 Observations of an uncorrelated multivariate gaussian 63
32 200 samples of a 100-dimensional gaussian distribution constructed as a linear transformation of the data plotted in Figure 31. Note the plaid pattern typical of low-rank matrices. 64
33 Eigenvalues of a Covariance Matrix 64
34 Subspace basis estimation on simulated data 66

35 Spectrogram plots of the spectral subtraction method 72
36 Spectrogram plots of the Wiener filter method 75
37 Spectral Subtraction before and after ISTFT 77

38 Spectrogram plots of the rank-1 filtering method 82

39 Signal Enhancement SDR Results 84

40 Rendering of the target location in the demo application. 90

41 Resynthesis Comparison 92
42 Rendering of a microphone in the demo application. 92

Acknowledgments

I am eternally grateful to the many people who have walked with me on this journey.

To my advisor, Joe Paradiso, for always putting your students first. Your genuine enthusiasm for ideas and deep respect for technical chops are deeply ingrained in the culture of Responsive Environments. I'm grateful for our conversations and for your continued support for my explorations.

To Dan Ellis, for giving me my first real glimpse of what research could be like, and being with me at the critical points ever since. I sincerely appreciate all the time you spent over the past year walking through my research with me, and nudging me in the right directions. During our conversations I always felt that you were listening and engaged, and your questions and suggestions were always insightful.

To Josh McDermott, for bringing a rigorous perspective while going along with my left-field projects, asking great questions, and always treating me and my work with respect.

To Brian Mayton, Ishwarya Ananthabhotla, and David Ramsay in the Responsive Environments Audio Crew. Our check-ins kept me on track when I needed it most, and I'm grateful for your questions, ideas, and critical role of pulling me out of rabbit holes.

The Tidmarsh project has been exceptionally collaborative, and I'm grateful to be a part of it. Thanks to Brian Mayton, Gershon Dublon, DDH, Clément Duhart, Félix Michaud, and all the others who have fought off angry, mosquitoes, thorny brambles, and leaky waders to help me fix microphones and play strange sounds from speakers. Also thanks to Glorianna Davenport, who had the vision to see with Tidmarsh could become, and continues to push the Living Observatory to new heights and depths. Without your constant drive to expand and innovate this work would not exist.

To Amna Carreiro, who is always there, making sure we are taking care of ourselves.

To Tod Machover and the rest of my friends in the Opera of the Future Group, for being a constant reminder of why I love music,

and why I love people who love music.

To the tireless work of the MAS Office, particularly Keira Horowitz, Linda Peterson, Monica Orta, Amanda Stoll, and Lily Zhang, who have been with me the whole time. You make sure we don't run afoul of the greater Institute, and obviously care deeply that everyone is safe and supported. Your mix of generous personal interactions and effective institutional change are an invaluable part of the Media Lab / MAS Department.

To all my friends on StudCom - past, present, and future. You are some of my favorite people in the Lab, and are always working to make it a better place for everyone.

To James Traer for your advice and example for capturing impulse responses in the field.

To my iZo DSP friends, particularly Gordon, Alexey, Hannah, and Russell - I've learned so much from you, and you're a great reason to stay in Boston.

To Fredrik Bagge Carlson for our conversations on low-rank techniques and opening my eyes to other perspectives.

To Zdeněk Průša for teaching me about frames and the STFT. I hope your pony is doing well.

To Vincent Lostanlen for our conversations on wildlife monitoring and time-frequency representations, among other things. Your kindness, enthusiasm, and thoughtful intelligence are infectious.

To Dale Joachim for our travels and helpful conversations about outdoor microphone arrays.

To Dan Gauger for your love of sound, and that S1 speaker.

To the Charlaaksos, Clafleets, Herbancos, and Slecksteins, so many babies. Thanks for putting up with me this past year or so, I can't wait to be a better friend.

To my Alyssa, Louis, Jenna, Steph, Angela, and Travis, who met me during a weird time in my life and have been critical in getting through it.

To Nick and Dave for sticking with me through many phases and always telling me what I need to hear.

To my SESI Friends, particularly Maya, Daniel, and Nick - as well as being great personal friends, you keep me connected to a part of myself that I sometimes neglect to my detriment.

To Bill, Jan, Lucas, and Z, without whom this dissertation literally would not have happened. I'll always appreciate your support and encouragement, and look forward to spending more time at the Chatham Compound next time I have a big paper to write.

To Karl and Ginny Birky, who have been a huge part of my life since the beginning. You are boundless reserves of love, and have

shaped who I am today.

To my Mom and Dad, the most encouraging people I know.

To Kate, without whom I never would have started or stopped this thing, and Frankie, who is always running.

Notation

I have relied on standard notation where possible, but notation varies from field to field. Given that this research spans multiple fields with sometimes conflicting notation, and in the interest of ensuring consistency within this document, I define some terms and notation here that is used throughout the dissertation.

Functions (interchangeably called signals) are considered to be mathematical objects that we can perform arithmetic on, and which can be evaluated at a given value with the common $f(t)$ notation. Unless otherwise noted or clear from context, signals are assumed to be complex-valued (thus results should generally also apply to real-valued signals as a special case). Both continuous and discrete signals act as vectors in the linear algebra sense. Generally arithmetic operators can be applied to functions, for instance $(f + g)(t)$ is equivalent to $f(t) + g(t)$. We do not use the relatively-common convention of using e.g. $f(t)$ to refer to a function - $f(t)$ is the result of evaluating f at t.

We also frequently treat random processes as signals. Performing arithmetic and evaluation adds an additional level of indirection - If x and y are random processes, $z = x + y$ gives another random process, and sampling from z is equivalent to sampling from x and y and adding the result. i.e. $z_0 \sim (x + y)$ is the same as $x_0 + y_0$ for $x_0 \sim x, y_0 \sim y$. Evaluating a distribution as $x(t)$ gives distribution over the results of evaluating samples of x. We also use standard notation for various properties of distributions, such as the expected value $\mathbb{E}[y]$ which produces a signal, or $\mathbb{E}[y(t)]$, which gives the expected value of the process y at time t. Note that if it is unclear which variable the expectation is being taken over, it can be given explicitly, for example $\mathbb{E}_i[v(i)]$ is the expected value of evaluating the signal v at a random point i.

We also define notation for a number of useful operators on functions, all of which are linear, and apply in both discrete and continuous domains:

$\mathcal{F}x$ Fourier Transform of x

$\mathcal{F}^{-1}\widetilde{x}$ Inverse Fourier Transform of $\widetilde{x}$. (We commonly will use $\widetilde{f}$ to represent the Fourier transform of f).

$\mathcal{R}x$ Reversal of x (reflection across zero), i.e. $\mathcal{R}x(t) = x(-t)$

$\mathcal{D}x$ Creates a linear operator (a matrix in the discrete case) with x along the main diagonal. This is particularly useful to represent elementwise multiplication, i.e. $Dxy(t) = x(t) * y(t)$.

These operators bind more tightly than function application, so $\mathcal{F}f(\omega)$ is evaluating the Fourier transform at the frequency ω, not taking the Fourier transform of $f(\omega)$, (recall $f(\omega)$ is a value in the range of f, not a function).

The scaling convention used in $\mathcal{F}$ and $\mathcal{F}^{-1}$ is defined such that Parseval's identity holds.

The *adjoint* of x is represented as $\overline{x}$. For scalars this is a complex conjugate, and for vectors and matrices it is the hermetian transpose.

$\|x\|$ is the L2 norm of x.

Introduction

Sound and space are fundamentally intertwined, at both a physical and perceptual level. Sound radiates from vibrating materials, filling space and creating a continuous field through which a listener moves. The technology to capture these sounds in space however, is currently limited to one of two approaches. One is an egocentric single-perspective recording, such as with binaural (or more recently ambisonic) microphones. The other is a labor- and expertise-intensive process of placing individual microphones on sources and adding an ad-hoc collection of general ambience microphones. This leaves open the question of how the recordings are related spatially, which requires yet more design and engineering.

In this work I propose and implement a new approach, where a distributed collection of microphones captures sound and space together, resynthesizing them for a (now-virtual) listener as a rich volumetric soundscape. The soundscape is constructed by decomposing the acoustic scene into its constituent components, collectively *auditory objects*. This approach provides great flexibility to design new auditory experiences, as well as giving a much more semantically-meaningful description of the space. The concept of the soundscape has been widely explored, most notably starting with Shaffer [1], and more recently in an analysis synthesis context [2], which also provides an excellent example of the labor required for effective volumetric soundscape capture. While this prior work covers the concept of soundscape (and more broadly acoustic ecology), from a perceptual and cultural context, I focus here primarily on the technical methods and tools necessary to capture and analyze such soundscapes.

Since 2013 my colleagues and I have been exploring various facets of technologically-mediated perception, with applications in remote presence as well as sensory augmentation. This work was articulated by Dublon and Paradiso [3] as tapping into environmental sensors as an extension of the human perceptual system. More fundamentally we've been exploring what it means to really

[1] R. Murray Schafer (1977). *The Tuning of the World*

[2] Andrea Valle, Vincenzo Lombardo, and Mattia Schirosa (2009). "Simulating the Soundscape through an Analysis/Resynthesis Methodology". Copenhagen, Denmark

[3] Gershon Dublon and Joseph A. Paradiso (July 2014). "Extra Sensory Perception: How a World Filled with Sensors Will Change the Way We See, Hear, Think and Live"

be somewhere at all.

Most of this work has been site-specific, and in a wetland restoration context. The Tidmarsh Wildlife Sanctuary is a 600-acre former cranberry farm that underwent the largest-ever freshwater restoration in the northeast. Since 2012 our group has instrumented the site with environmental sensors (e.g. temperature, humidity, soil moisture), microphones such as the one depicted in Figure 2, and cameras [4]. We are collaborating with a variety of other researchers from other institutions that are investigating ecological questions related to the restoration.

We've built a number of interfaces for users to interact with Tidmarsh remotely such as Hakoniwa[5], and Doppelmarsh[6] [7]. These interfaces provide an augmented-reality (AR) and virtual-reality (VR) experience respectively. They allow users to engage with different aspects of Tidmarsh from afar. Both systems bring the live microphone streams into the virtual environment, and also generate music that responds to the real-time sensor data [8]. We've also worked on the on-site experience in the HearThere project [9] which uses head-tracking bone-conduction headphones to let a user tap into far-away sounds or sonified sensor data while they are at the marsh. This is one of the most direct instantiations of the sensors-as-human-perception vision.

While these projects provided experience and insight into the listener's experience of Tidmarsh, the lack of a clear mapping between what the microphones captured and what the listener should hear was a recurring issue. This was the initial motivation for the work in this dissertation. The high-level goal is to combine the signals from each microphone and resynthesize a plausible and meaningful auditory scene at an arbitrary listening position. The microphones are scattered across an area of the marsh roughly 300 by 300 m^2, which includes both open marsh and a forested area, as seen in Figure 1.

The framework I am adopting treats this as a problem of simultaneous localization and source separation. The goal of source separation is to extract a target sound from a mixture of noise and interfering sounds. The separated foreground signal such as a bird call is combined with its location to create a virtual sound source that can be rendered from the perspective of an arbitrary listener position. The residual background noise at each microphone is used to create a diffuse background soundfield. An off-the-shelf audio engine is responsible for rendering the auditory scene to the listener.

This leads to the question of what defines a foreground sound. In this work I consider foreground sounds to be those that are

[4] Brian Mayton et al. (May 2017). "The Networked Sensory Landscape: Capturing and Experiencing Ecological Change Across Scales"

[5] `https://vimeo.com/212681207`

[6] `https://vimeo.com/240549912`

[7] Don Derek Haddad et al. (2017). "Resynthesizing Reality: Driving Vivid Virtual Environments from Sensor Networks"

[8] Evan F Lynch and Joseph A Paradiso (2016). "SensorChimes: Musical Mapping for Sensor Networks". Brisbane, Australia

[9] Spencer Russell, Gershon Dublon, and Joseph A. Paradiso (2016). "HearThere: Networked Sensory Prosthetics Through Auditory Augmented Reality"; Gershon Dublon (2018). "Sensor(y) Landscapes:Technologies for New Perceptual Sensibilities". Doctoral Dissertation. Cambridge, MA

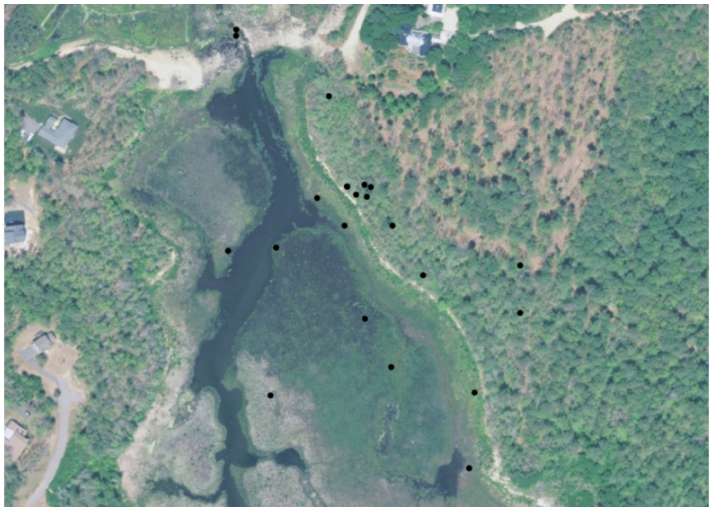

Figure 1: A map of Tidmarsh showing our microphone locations. The mapped area is about 500 by $350m^2$.

Figure 2: A sampling of equipment deployed at Tidmarsh. (Clockwise from upper left) Some microphones ready for deployment, a microphone in a tree in the forested area, a 32-channel USB audio interface with attached encoding computer, a microphone on a pole in the marsh, an expansion interface providing an additional 16 channels which is connected to the primary interface via Cat5 cable. (Figure used with permission from Mayton et. al)

spatially-compact (they can be modeled as a point source). This is obviously not the only definition of foreground sounds, and not necessarily the most appropriate for all circumstances, but I argue that it is very useful in the soundscape resynthesis context. The purpose of segregating the foreground sounds is to be able to provide more accurate spatial cues to a listener as to the sounds location. Thus for sounds that are not spatially compact (such as wind, or a large flock of birds) it is less important to accurately spatialize them, and they can be treated as diffuse background sound.

Problem Statement

Our VR and AR applications allow the user to move arbitrarily around the monitored area, yet we only measure the soundscape at the individual microphone locations. For a given listener location, we need a way to map point measurements onto a stereo signal that the listener can hear through headphones. Further, this rendering should reflect the relative locations of listener and source. The first approach we tried was to create a source at the location of each microphone, a sort of virtual speaker. This is straightforward to do with commercially-available game audio engines. These engines can spatialize such sources into headphones relative to the listener's virtual head position, providing spatial perception. They also provide control over the effective area that a source appears to emit from, permitting either point sources or diffuse sources.

This approach has several problems. For sounds only audible in a single microphone, this approach has the effect of localizing the sound at the location of the microphone instead of the source. For sounds audible through multiple microphones the situation is even worse - delayed copies of the sound from each microphone create a kind of "virtual multipath", perceived by the listener as echos or reflections.

In this work I propose a more informative approach: given a set of microphone signals at known locations, the system uses all of them together to estimate a *meaningful* and *convincing* 3D soundscape that can be experienced at arbitrary locations. Meaningful indicates that result should provide a listener with a perception of what is actually happening on the site. Convincing indicates that the listener should have a sense that what they're hearing could plausibly be what they would hear were they actually present on-site.

These criteria were chosen in part because they provide a useful

tension and keep the work focused on the experience of a listener. For instance, a verbal description of the sounds on the site and their locations is a meaningful representation, but is not connected to the perceptual experience of being on site. Conversely, a completely synthetic but carefully-designed soundscape could be very convincing but is disconnected from reality.

Applications

Outside of the VR and AR context that initially motivated this work there are a number of other applications in wildlife monitoring and conservation. The proposed approach, when combined with a classification system [10] would allow detailed monitoring of the movements and micro-habitats of different species. While there is an abundance of literature on acoustic localization for wildlife, it is typically performed over short durations (on the order of hours), and often with a specific target in mind that is manually annotated [11]. Others, such as the HARK system [12] perform separation and direction-of-arrival finding, but only support small arrays with sources in the farfield.

Modern source separation techniques using deep learning have enabled new tools for music production [13] [14] but so far spatial information (if used at all) has been modeled either with instantaneous mixtures, or within the framework of farfield arrays, where the inter-microphone delay is small compared to the wavelengths of interest. The proposed approach could be applied to multichannel recordings for the purposes of bleed (inter-microphone interference) reduction, or remixing.

It could also enable new recording techniques where an area is instrumented with a spaced array of microphones rather than recording each source with its own microphone. A spatially-aware representation would also enable new types of analysis of a space, either for the purposes of capturing its "essence" and resynthesizing an infinite non-looping soundscape, or for compression and prioritization. These representations could be particularly rich when combined with the perceptual and semantic features that affect how a listener hears and remembers sounds [15]. Object-based representations such as Dolby Atmos have recently been gaining traction in the gaming and home theater spaces, but the tools to capture a space in these formats have been lagging.

This type of array could also be applied to conferencing systems, or factory monitoring contexts where the sound produced by individual machines can give valuable information about its operating state.

[10] Clement Duhart et al. (Oct. 2019). "Deep Learning for Environmental Sensing Toward Social Wildlife Database". Paris, France; Stefan Kahl et al. (2019). "Overview of BirdCLEF 2019: Large-Scale Bird Recognition in Soundscapes"

[11] Daniel J. Mennill et al. (Apr. 2006). "Accuracy of an Acoustic Location System for Monitoring the Position of Duetting Songbirds in Tropical Forest"; Travis C. Collier, Alexander N. G. Kirschel, and Charles E. Taylor (July 2010). "Acoustic Localization of Antbirds in a Mexican Rainforest Using a Wireless Sensor Network"

[12] S. Sumitani et al. (May 2019). "An Integrated Framework for Field Recording, Localization, Classification and Annotation of Birdsongs Using Robot Audition Techniques — Harkbird 2.0"

[13] `https://web.archive.org/web/20200319181238/https://www.izotope.com/en/products/rx/features/music-rebalance.html`

[14] Romain Hennequin et al. (2019). "Spleeter: A Fast and State-of-the Art Music Source Separation Tool with Pre-Trained Models"

[15] Ishwarya Ananthabhotla, David B. Ramsay, and Joseph A. Paradiso (2019). "HCU400: An Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty"

Research Questions

With this context, we can further focus the research by working to answer the following questions:

What field measurements are necessary to support localization and separation?
Outdoor environments have complex acoustics due to irregular landscape variations and complex geometry. Simulating acoustic propagation at a particular site would require an impractically-accurate 3D model, so on-site measurements play an important role in system design and validation.

How do farfield beamforming DSP techniques extend to our nearfield and aliased condition?
Many sensor array techniques assume that the source is far from the array, so the incoming wave can be considered to be a plane wave. Additionally the spacing of the array is typically small relative to the signal wavelengths. This work explores the nearfield regime (the sources are within the sensor array) and in the presence of significant spatial aliasing. Which existing techniques are applicable and where do they need to be modified?

What elements are necessary to resynthesize a meaningful and convincing spatial soundscape?
This work is rooted in the perceived experience of a listener. Rather than attempting to solve the ill-posed question of minimizing actual reconstruction error (e.g. in the mean-squared error sense), my approach is to elicit in the listener a percept that is meaningfully connected to the wildlife and ambient environment on-site. This work develops and implements a framework for soundfield resynthesis, demonstrated by an end-to-end application that resynthesizes a soundscape captured on site.

The remainder of this document is organized around these questions. Part I describes the microphone installation and other infrastructure installed at Tidmarsh, as well as a variety of acoustic measurements that were performed to characterize the site. The second research question is addressed in Parts II and III, which describe acoustic array algorithms for performing localization and separation, respectively. Finally, Part IV describes an interactive application that renders the separated sounds from the perspective of a listener.

Part I

Characterizing the Environment

Background: Impulse Response Measurement

The success of a localization and separation system is contingent on assumptions about acoustic propagation at the specific site being monitored. We performed a variety of measurements on site to characterize the propagation and validate or refute these assumptions. There are many techniques in the literature and practice to measure impulse responses. Most are functionally equivalent given the assumption that the system under test is noiseless, linear and time-invariant, but they perform differently when these assumptions are violated. I have implemented several approaches: exponential sine sweep [16], maximum-length sequences [17], Golay codes [18], random phase multisine [19], and simple dirac delta (for comparison).

The same general framework is used by all these methods:

1. Generate a stimulus. This stimulus should have energy in all the frequencies of interest.
2. Play the stimulus through the system under test, and record the response. We refer to this as the *stimulus response*.
3. Analyze the stimulus response to extract the underlying *impulse response*, which estimates what the system's response to a true impulse would be.

Assuming the system is time-invariant, the stimulus response will be periodic with the same period as the stimulus, so it is common to average multiple repetitions for to improve the effective signal-to-noise ratio. This is because the energy due to the stimulus will add coherently while the noise will not. Ideally analyzing the unmodified stimulus should result in a perfect dirac delta.

[16] Angelo Farina (2000). "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique"

[17] Martin Holters, Tobias Corbach, and Udo Zölzer (2009). "Impulse Response Measurement Techniques and Their Applicability in the Real World". Milan, Italy; Wayne Stahnke (1973). "Primitive Binary Polynomials"

[18] Edgar J. Berdahl and Julius O. Smith (June 2008). *Transfer Function Measurement Toolbox*; M. Golay (Apr. 1961). "Complementary Series"; S. Foster (Apr. 1986). "Impulse Response Measurement Using Golay Codes"

[19] I. Mateljan (1999). "Signal Selection for the Room Acoustics Measurement"

Dirac Delta

The simplest way to record an impulse response is to generate a signal that approximates an actual impulse. The recording can

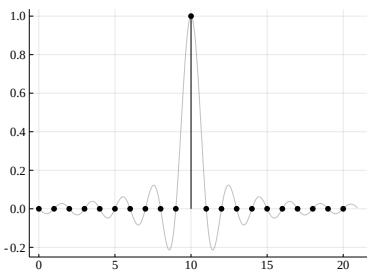

Figure 3: Time-domain plot of a dirac delta. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal.

then be used as-is without further processing. The main disadvantage of this method is that all the energy of the stimulus is concentrated in time, so must have very high amplitude in order to get an adequate signal-to-noise ratio. Sufficient instantaneous power is difficult to achieve with a speaker while remaining in its linear region, and even with an acoustic source such as a balloon pop or starter pistol, there is danger of driving the capture equipment into nonlinearity. Balloons and pistols also offer much less control of the stimulus spectrum. This limits their applicability for capturing impulse responses intended for simulation via convolution, though they can still be useful for characterizing reverberation statistics like decay time and direct-to-reverberant ratio.

One benefit of these true impulse approaches is that if the system is time-variant, they capture the actual response at the measurement time.

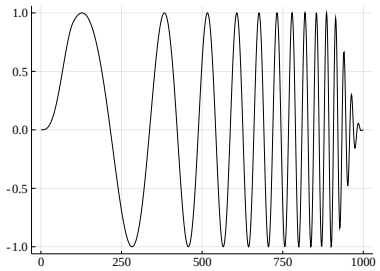

Figure 4: Time-domain plot of an exponential sine sweep.

Exponential Sine Sweep

The exponential sine sweep (ESS) method spreads the stimulus energy over time with a sinusoid whose frequency is modulated from low to high. The signal has energy in all the swept frequencies. The spectrum is not at all flat however, which must be compensated for when analyzing. An ESS response is analyzed by convolving with a time-reversed version of the stimulus, where the amplitude is modulated to compensate for the non-flat stimulus. The group delay of the analysis filter has the effect of shifting each frequency to turn the input sweep into an impulse. The primary benefit of ESS is that it separates out any nonlinear effects of the system. Because the sweep goes from low to high, any harmonics of the stimulus show up in the noncausal region of the impulse response, and can be cropped out, or used for nonlinear system modeling.

The main downside of the ESS method is that it is sensitive to impulsive noise during the measurement process. Because analysis consists of convolving the stimulus response with an inverse filter that's a downward frequency sweep, any impulsive sounds that occurred during the recording process will generate audible artifacts are a copy of the downward sweep. Thus this method is not very suitable if there is likely to be uncontrolled background noise in the system.

Another downside of ESS is that there are a large number of variations and parameters that need to be selected, such as start/stop frequencies, fade times to prevent discontinuities, and

the exact curve of the frequency-modulating exponential. On the other hand, this makes it much easier to control the bandwidth of the stimulus, and thus the impulse response.

Maximum-Length Sequence

The Maximum-Length Sequence method (MLS, also called Schroeder's Method) uses a linear-feedback shift register similar to what is often used by pseudo-random number generators, which can be done very efficiently (though on modern computers this is less of a consideration, particularly when the stimulus is generated once and re-used many times). Analysis is performed via circular cross-correlation, though because the stimulus signal is ±1, the cross-correlation does not require any multiplications, only sign-flips and addition [20]. Because each sample of the signal is ±1, it has maximum digital "energy" for a given maximum amplitude (low crest factor). However, if you consider the digital signal as samples from a bandlimited continuous-time signal, the amplitude is considerably larger than the sample values because of inter-sample peaks, which must be accounted for when calibrating playback equipment.

These sequences have the property that their circular autocorrelation has value 1 when the offset is 0, and $-\frac{1}{L}$ otherwise (where L is the length of the sequence). This slight offset is because the MLS sequence does not have a perfectly flat spectrum, but has an attenuated DC component. At lengths typically used in acoustics this offset is generally insignificant, and DC properties are not generally relevant in acoustical systems.

Because this requires circular cross-correlation, typically the stimulus consists of two or more repetitions of the sequence.

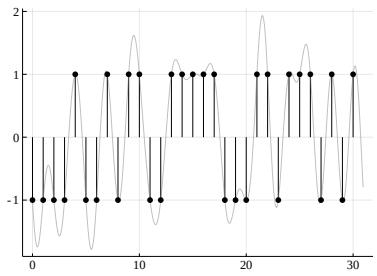

Figure 5: Time-domain plot of a maximum-length sequence. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal.

[20] Martin Holters, Tobias Corbach, and Udo Zölzer (2009). "Impulse Response Measurement Techniques and Their Applicability in the Real World". Milan, Italy

Random-Phase Multisine

The Random-Phase Multisine (RPMS) method generates a broadband stimulus in the frequency domain by synthesizing a flat magnitude spectrum and randomizing the phase, then generating the time-domain signal with the inverse FFT. This is easy to do, and the analysis is a simple circular cross-correlation with the stimulus signal. Additionally because the energy is spread across all frequencies at all times it does not have the same artifacts as the ESS method in the presence of impulsive noise. Like the Maximum-Length Sequence approach, the response is analyzed by performing circular cross-correlation, which for an identity system

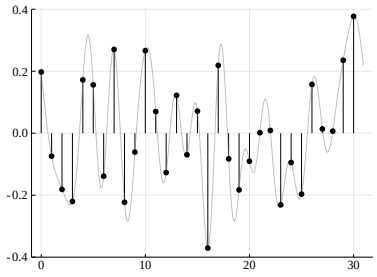

Figure 6: Time-domain plot of a random-phase multisine stimulus. The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal.

will recover a perfect impulse. Performing the cross-correlation (generally performed in the frequency domain) is more costly than MLS, which maybe relevant for embedded systems. However, RPMS has the advantage that it is easy to design the spectrum of the stimulus to match the needs of the measurement (e.g. using "pink" noise rather than white). Shaping the spectrum of MLS would require filtering the stimulus and response, which negates its performance benefits.

Complementary Golay Sequence

Golay sequences involve generating two stimuli, generally called A and B, which have the property that the sum of their autocorrelations is a dirac delta. This means that by doing our stimulus/analysis with each sequence separately, and then adding the results, we get a perfect impulse measurement. One of the advantages of this over MLS and RPMS is that it works under linear convolution, so repeated stimuli are not necessary.

However, this approach is somewhat more sensitive to clock skew errors because it relies on the alignment cancellation between A and B when recombining their responses. This has been observed in the context of HRTF measurement [21], where it was shown to be more sensitive to time-variance.

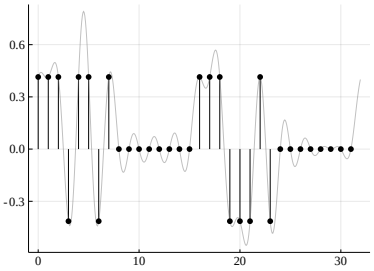

Figure 7: Time-domain plot of a Golay sequence (parts A and B). The black dots indicate digital samples and the grey line traces the corresponding continuous-time signal.

[21] Pavel Zahorik (Feb. 2000). "Limitations in Using Golay Codes for Head-Related Transfer Function Measurement"

Audio Infrastructure Overview

The Responsive Environments Group has been maintaining a microphone installation at Tidmarsh in one form or another since 2013, and we have learned a number of lessons over the duration of the deployment. Equipment has been damaged by rodents chewing on the cables, as well as construction work and lightning strikes, and we have iterated substantially on the deployment. This part of the work has been particularly collaborative. The initial deployment was performed by Gershon Dublon and Brian Mayton [22], and since Spring 2018 it has been maintained and extended by Brian Mayton and myself. Many aspects of the installation will be described in more detail in Mayton's upcoming PhD dissertation. I designed and led the field experiments, with assistance from Brian Mayton and many others in the Responsive Environments group and the volunteers from the Living Observatory (a collaboration between scientists, artists, and wetland restoration practitioners).

[22] Gershon Dublon (2018). "Sensor(y) Landscapes:Technologies for New Perceptual Sensibilities". Doctoral Dissertation. Cambridge, MA

This chapter will describe the microphone array deployed at Tidmarsh, as well as a series of field experiments we performed. The experiments were designed to characterize the acoustic propagation (for which we captured impulse responses at various locations), as well as record ground truth audio at known locations that could be used to test the localization and separation algorithms.

The on-site recording system includes between 10 and 18 microphones (the number has varied over time). These are connected via cables to an audio interface (Behringer X32 RACK) connected to a small PC (Intel NUC running GNU/Linux). The PC compresses the data and streams it to the lab, where it is archived. Because the microphones are connected to a single audio interface, they are synchronized, enabling accurate measurements of the time delays between the arrival of a sound at different channels.

The microphones are a custom design by Brian Mayton and consist of:

- EM272J Omnidirectional electret condenser microphone capsule

from Primo Microphones Inc.
- Custom buffer circuit
- Aluminum enclosure (pipe cut to length)
- "Pigtail" cable with Switchcraft EN3C3MX waterproof 3-terminal connector
- Reynolds OOMOO 30 silicone rubber

The capsule, buffer, and cable are soldered together and the capsule and buffer are placed within the aluminum tube. The tube is filled with silicone, with a jig ensuring that the capsules remain exposed.

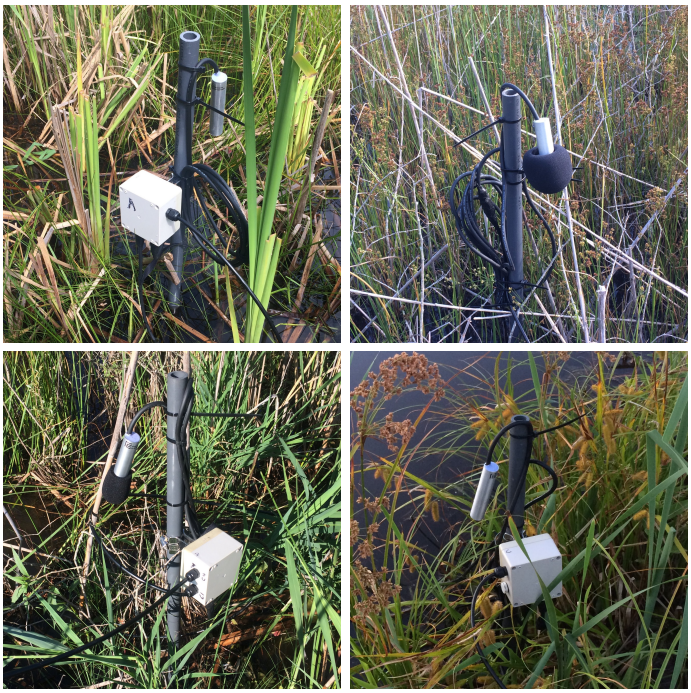

Figure 8: Four microphones deployed at Tidmarsh. The boxes visible on the poles include a waterproof panel-mount connector for the microphone, as well as cable glands to seal the cable coming in, as well as the cables that continue on to other microphones in a daisy-chain.

The initial installation used standard microphone cable to connect each microphone to the audio interface in a "home-run" configuration. The cabling was submerged throughout most of the wetland area, and lightly covered with leaves and brush where above ground, and in the forested area. The most frequent issue we encountered was damage from rodents chewing on the cables. Also the cables were difficult to splice in the field, as all the conductors needed to be soldered and the splice sealed. In 2018 we replaced most of the cables with 24AWG gel-filled direct-burial shielded cat5e cable. The cat5e cable has higher series resistance (19Ω/100m vs 13Ω/100m for high-quality mic cable), but lower capacitance (52pF/m vs 162pF/m) and is designed to manage

much wider bandwidth than what we find in audio. This has a number of advantages.

- The sheath of the cable is designed for long-term outdoor use, and seems less susceptible to rodent chewing, or simply less attractive to rodents.
- In the case the cable is pierced, the gel filling prevents water from wicking further down the cable by capillary action.
- We are able to use the four pairs within the cat5e cable to carry balanced audio for four channels, allowing us to "daisy-chain" the microphones rather than needing a home-run for each microphone.
- There are a large number of tools developed by the telecommunications industry that we are able to leverage to ease maintenance. For example, splice boxes are available with waterproof glands and small punch-down blocks inside, so only the shield wire needs to be soldered. We have also been able to use standard punch-down blocks for making connections inside the larger boxes, enabling greater flexibility and ease of maintenance.
- The controlled impedance and high bandwidth (100MHz vs. 50kHz) enables the use of a time-delay reflectometer, which can identify the location of cable faults by sending a sharp pulse along the able and measuring the reflection emitted by faults.
- When cable damage is limited to one pair of conductors and another pair is available, we can switch the microphone connections at the ends of the cable without requiring a splice.

Source localization depends on accurate knowledge of the microphone locations. We used a real-time kinematic (RTK) GPS system developed by Brian Mayton, which measures local delays in GPS signal at a stationary GPS radio with known location. These delays are caused by distortions in the atmosphere, and are time-varying. Measurements from the stationary GPS radio are relayed in real-time to the mobile GPS device, where they are used to apply correction factors, permitting location accuracy to within 0.5m.

Speaker Preparation

Field recordings were performed using a Bose S1 portable speaker, with recordings stored on and played from an iPhone XR into the speaker's analog input. To prepare for data collection, I performed preliminary tests, primarily to validate speaker linearity, and capture calibration responses that could be used to compensate for the transfer function of the speaker itself (though I did not end up making use of these calibration responses).

Measurements were performed in the Multipurpose room on the 6th floor of building E14 at the MIT Media Lab, which is a large room (18.6m by 18.2m) with carpeted floor. The speaker and microphone were placed 2.1m" off the ground and 1.8m from each other (measured speaker face to mic capsule). This should allow the measured impulse response to be truncated to remove the effects of the wall reflections, though reflections from the floor would not be able to be removed without sacrificing information about the low-frequency response of the speaker.

Figure 9: Speaker and microphone used for speaker measurements

The first test was to prepare Maximum-length-sequence (MLS) stimuli at a range of gain levels, and to verify that the measured sound levels were as expected. This was primarily to validate that our audio pipeline was not introducing any automatic gain control or limiting that would affect our experiments. All hardware gains were kept constant, and the signal gain was varied by scaling the audio files used for playback. Measurements were performed using a handheld SPL meter placed next to the microphone capsule.

From Figure 10 we observe an ambient noise floor of about 56dB SPL (C-weighted). Once the gain was sufficient that the energy at the meter was dominated by the speaker output, we see the output increasing as expected.

I then performed exponential sweeps to characterize the system's linearity. Figure 11 shows four of the resulting recordings. We see that with an 80dB range displayed, there is visible distortion with a sample gain of about -25dB. In this case the reference level was a full-scale (0dBFS) sine sweep, though the absolute level is not very meaningful because the total output is also affected

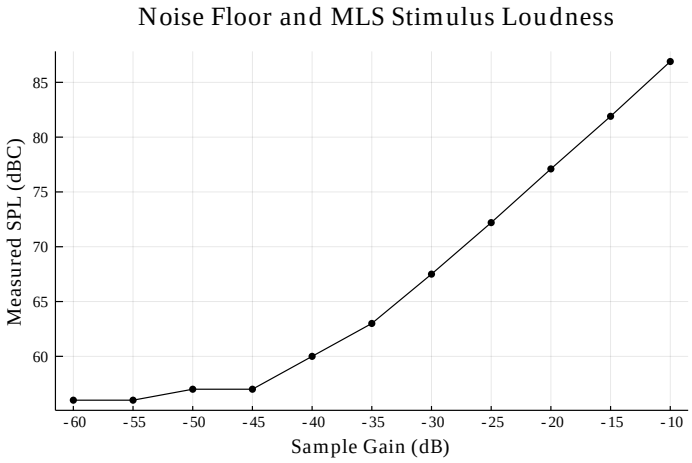

Figure 10: Speaker output SPL against sample gain

by the playback device gain as well as the speaker input gain. Device and speaker gain levels were recorded so that signals could be generated with confidence they would not be distorted. This information was used to set a maximum amplitude used for the following field experiments, which were set to keep the amplitude under -30dBFS.

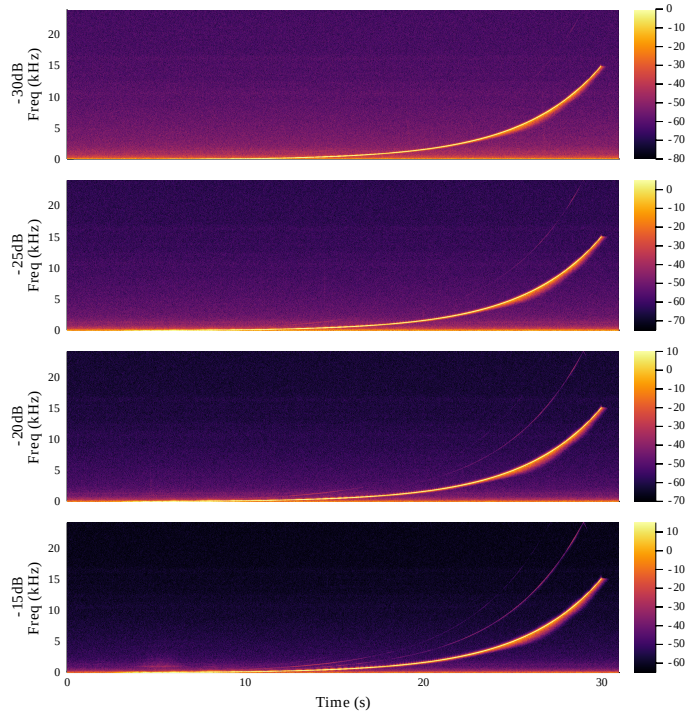

Figure 11: Spectrograms of exponential sine sweep recordings at different gain levels. Each shows an 80dB range, shifted to account for the gain of each recording, to emphasize the relative power of the distortion signal vs. the main signal.

Impulse Response Capture

Acoustic propagation on the site was characterized by measuring impulse responses from a variety of locations, recording on all microphone channels simultaneously. The measurements were taken by two means - the simplest was a slapstick (a percussion instrument made from two wooden planks connected with a spring and hinge), which was used to generate impulses. Additionally stimuli were prepared using maximum-length sequences (MLS). See Part II for more background comparing different methods of measuring impulse responses.

Measurements were performed with a Bose S1 portable speaker. The stimulus recordings were played from an iPhone XR, via the speaker's analog input.

Visual inspections of preliminary recordings using the slapstick indicated that the decay time was around 0.75s, so MLS stimuli were generated with a period of 65535 samples (1.36s at 48kHz sampling rate). The stimulus contained 80 repetitions, for a total length of 109s).

Each MLS stimulus was preceded by a 10s 1kHz pilot tone to provide a synchronization reference.

Figure 12: Speaker set up in the field

Clock Skew Compensation

Because of the large scale and dense foliage present at the field site, it is impractical to use a cable between the playback speaker to the system recording from the microphones. However, if the playback and recording devices are running at slightly different clock frequencies (due to crystal manufacturing variance, temperature, etc.), it will appear as if the response has been slightly slowed down or sped up (time dilation). This is known as "clock skew" [23] (also "clock drift" or "frequency offset" [24] in some contexts).

It's difficult to get specifications for clock accuracy on many audio devices. Some high-end devices like the Tascam CG-2000[25] report an accuracy of 0.01 ppm (parts per million), which provides a rough upper bound for pro audio clock accuracy. Professional audio interfaces have also been observed to vary by 15-30ppm [26]. Most handheld consumer devices are accurate on the order of 10ppm, with some outliers up to 500ppm [27]. This gives a worst-case relative drift of 1000ppm.

Time variance in the context of impulse responses can be characterized as having two different effects, described as "within-frame" and "between-frame" effects [28]. Clock skew between the playback and capture systems causes both types. Assume that we are extracting the impulse response from the stimulus response by means of convolution with some kind of matching filter that acts as the inverse of the stimulus. The first effect of clock skew is that the filter no longer matches the stimulus signal - it is slightly stretched or compressed in time. This will tend to cause spreading in the resulting impulse response. The second effect is due to the synchronous averaging typically performed when measuring impulse responses. If the response is assumed to be periodic with period P, but has true period P', each frame will have a delay of $P' - P$ relative to the previous one. When the frames are summed, this will in effect create a comb filter. These effects, and methods to compensate for them have been investigated in the context of sine-sweep measurements [29], where the clock drift can be modeled as an allpass filter with the appropriate group delays.

[23] D. Mills (Mar. 1992). *Network Time Protocol (Version 3) Specification, Implementation and Analysis*. Tech. rep.

[24] Michael A. Lombardi (2010). "Time and Frequency from A to Z"

[25] `https://tascam.com/us/product/cg-2000/`

[26] Nicholas J. Bryan, Miriam A. Kolar, and Jonathan S. Abel (2010). "Impulse Response Measurements in the Presence of Clock Drift"

[27] Mario Guggenberger, Mathias Lux, and Laszlo Böszörmenyi (2015). "An Analysis of Time Drift in Hand-Held Recording Devices"

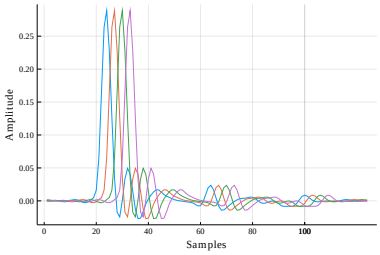

Figure 13: Multiple skewed frames superimposed. Summing them together creates a comb filtering effect

[28] Peter Svensson and Johan L. Nielsen (May 1996). "Errors in MLS Measurements Caused by Time Variance in Acoustic Systems"

[29] Hannes Gamper (2017). "Clock Drift Estimation and Compensation for Asynchronous Impulse Response Measurements". San Francisco, CA, USA; Nicholas J. Bryan, Miriam A. Kolar, and Jonathan S. Abel (2010). "Impulse Response Measurements in the Presence of Clock Drift"

Bryan et. al [30] propose a skew estimator using an impulse or chirp train, where the skew is estimated by recording through the system, performing peak picking, and measuring the time differences between the peaks. Gamper observed that the stimulus itself could be used for skew estimation by estimating the delay of each frame, minimizing the objective function given by Equation 1 (adapted to be in discrete rather than continuous terms). The response is first sliced into R frames of length P, corresponding to each repetition of its nominal period.

[30] Nicholas J. Bryan, Miriam A. Kolar, and Jonathan S. Abel (2010). "Impulse Response Measurements in the Presence of Clock Drift"

$$\hat{d} = \underset{d}{\operatorname{argmin}} \sum_{r=1}^{R-1} \sum_{k=0}^{N-1} \left| \widetilde{x}_0(k) - \exp\left(2\pi j d r \frac{k}{N}\right) \widetilde{x}_r(k) \right|^2 \qquad (1)$$

Here $\widetilde{x}_r$ is the discrete Fourier transform of the rth frame of the response (assuming the nominal period), N is the FFT size (larger than P). This objective function compares each frame to the first one, shifting the rth frame by dr samples. Performing this operation in the frequency domain enables sub-sample delays to be found, as the complex exponential term represents the time delay as a linear phase shift.

In practice, this function has three main issues. First, this function gives the first frame special status as the reference. This is not necessarily justified, particularly in a noisy environment where any particular frame might be corrupted by noise. Secondly, for measurements with a large number of frames, small variations in d result in large shifts in the later frames (r is large). This gives greater weight to later frames. The third issue is a practical concern - in our context the sensitivity to small variations in d resulted in an objective function that was difficult to optimize - for the analyzed data it had a very narrow trough surrounding the true minimum, with many nearby local minima.

Rather than comparing each frame to the first, we propose comparing each frame to the next frame. This is equivalent to comparing the entire signal with itself, shifted by one nominal period.

$$\hat{d} = \underset{d}{\operatorname{argmin}} \sum_{k=0}^{N-1} \left| \widetilde{x}(k) - \exp\left(2\pi j (d+P) \frac{k}{N}\right) \widetilde{x}(k) \right|^2 \qquad (2)$$

As above, N is the the FFT size, though now we are transforming the entire signal rather than working with individual frames. This has a performance consideration because the FFT and IFFT scale at $\mathcal{O}(n \log n)$ where n is the total signal length, whereas the Gamper objective function performs the FFT on fixed-size blocks so it is $\mathcal{O}(n)$ in the total signal length. The proposed objective function could be approximated at the same cost as Gamper by

performing the FFT in each block and comparing each block to the next, at the expense of some additional framing effects.

Recall that P is the nominal period. If we define $\widetilde{x}_d(k) = \exp\left(2\pi j(d+P)\frac{k}{N}\right)\widetilde{x}(k)$, we can then define the time domain versions of $\widetilde{x}$ and $\widetilde{x}_d$ as x and x_d respectively. So our objective function can be written as

$$e(d) = \sum_{k=0}^{N-1} |\widetilde{x}(k) - \widetilde{x}_d(k)|^2$$

Invoking Parseval's identity, the energy of a signal is the same in the time and frequency domains, so this is equal to

$$e(d) = \sum_{n=0}^{N-1} \left|\mathcal{F}^{-1}(\widetilde{x} - \widetilde{x}_d)(n)\right|^2$$

Because the Fourier transform is linear this becomes:

$$e(d) = \sum_{n=0}^{N-1} \left|\mathcal{F}^{-1}\widetilde{x}(n) - \mathcal{F}^{-1}\widetilde{x}_d(n)\right|^2 = \sum_{n=0}^{N-1} |x(n) - x_d(n)|^2 \qquad (3)$$

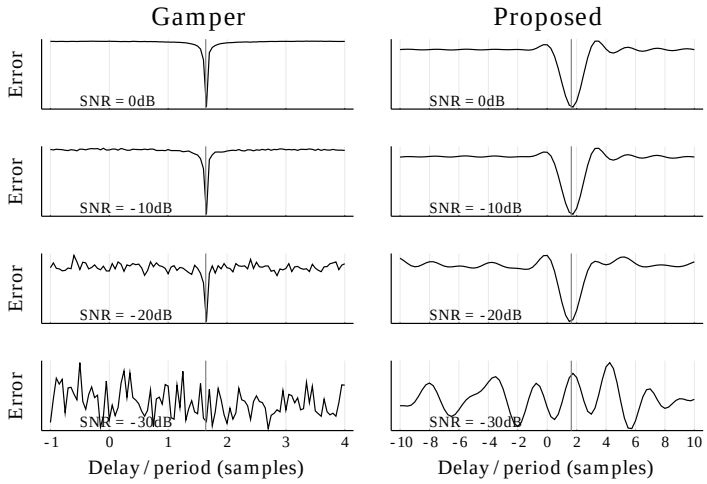

Figure 14: This shows the objective function to be optimized to determine the period of a clock-skewed impulse response. The delay per period (X axis) shows how much each frame is delayed relative to its expected position given the nominal period. These plots were generated with simulated MLS data with 80 repetitions and period 65535. The response was skewed by 25ppm via resampling, correspondig to a per-period delay of 1.64 samples (shown by the grey bar in each plot).

So Equation 3 shows us that the proposed frequency-domain objective function is equivalent to the Average Squared Distance function (ADSF), which is widely used in time-delay estimation but to my knowledge has never been applied to clock skew estimation. The ASDF has been shown to be generally provide a more accurate estimated delay than the cross correlation [31], but the difference is generally less than one sample. $|x - x_d|^2$ is bandlimited to twice the bandwidth of x, which limits how sharp the trough containing the minimum can be. The algorithm is initialized by first performing a discrete autocorrelation, which can be

[31] Giovanni Jacovitti and Gaetano Scarano (1993). "Discrete Time Techniques for Time Delay Estimation"

efficiently computed via the FFT, or for a small number of lags can be computed directly in the time domain. The correlation is then up-sampled by 4 to detect inter-sample peaks, and the maximum in a window near P is chosen as an initial estimate for the period. This estimate should be within 0.25 samples of the true period, which can then be found with by simple Newton's Method optimization of Equation 2.

Figure 14 shows a comparison of these objective functions in a neighborhood of the true period, for different signal-to-noise ratios. In informal testing the narrower peak of the Gamper function does not lead to greater accuracy, though more investigation is necessary to determine the trade-offs more fully.

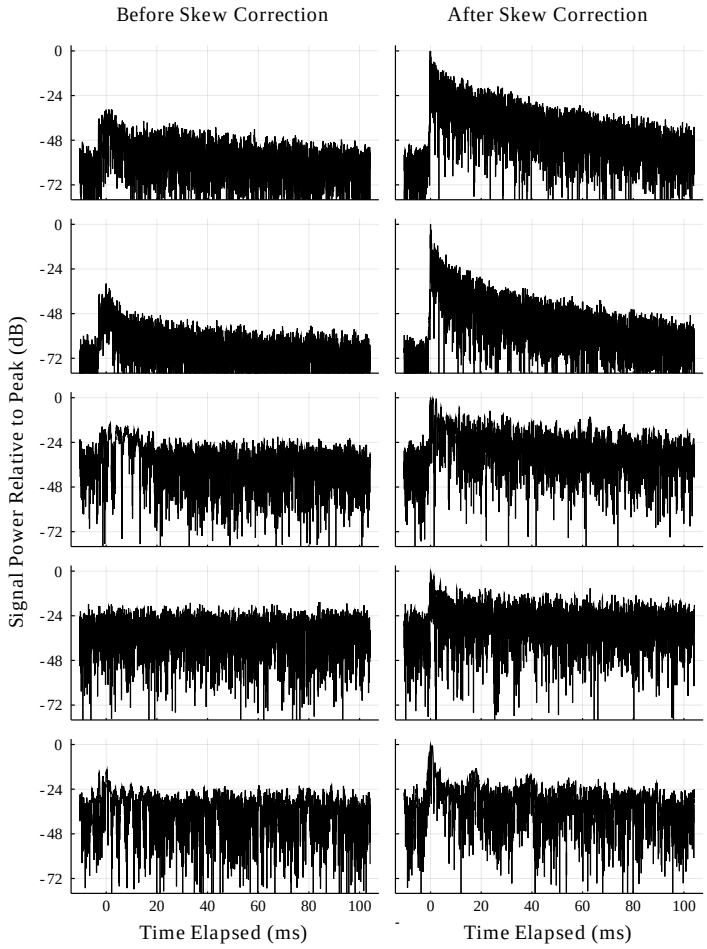

Figure 15: Clock skew correction applied to several stimulus responses. This shows the result of extracting the impulse response from the MLS stimulus response, with and without the clock skew correction. These examples are channels 2, 3, 10, 11, and 13 from source location 020-E.

Figure 15 shows some examples from our dataset of on-site MLS recordings, where the clock skew was estimated to be -26ppm during these recordings (note the skew is expected to drift with time and temperature). Notice that the peak power to noise

ratio is improved by 30dB in the best cases, but results vary and further characterization is needed. The degree of improvement is sensitive to variations in the skew estimate - in informal tests changing the skew by 2ppm caused the peaks to be 3-6dB smaller.

One possible improvement to the proposed skew estimator would be to incorporate that phase transform (PHAT). While at first it may not seem like this would provide much gain because the stimulus is broad-band, it would help reduce the impact of interfering signals with harmonic energy.

Additionally, we currently assume a constant clock skew for the duration of the recording. While this model works well to account for skew due to to the actual hardware crystal, modeling a time-varying skew may help account for other time-variances, such as changing wind speed or direction.

Results

Clock Skew Estimation

For each impulse response measurement, the clock skew was estimated using the process described in Part I. Figure 16 shows a histogram of the estimates, and displays a clear peak at the median value. No ground truth is available so it's impossible to report an absolute error, but from the examples shown in Part II we see that resampling the recording based on this skew estimate provides substantially improved impulse response signal-to-noise.

Figure 17 shows the results grouped by their location and speaker direction, displayed in chronological order of when they were recorded. All recordings were performed in a single day. We do not see any evidence of drift in the clock skew over the recording period, which is consistent with earlier work [32] that reported relatively constant clock skews over a five-hour period, but significant variations day-to-day. For environments with more severe temperature variations the effect of temperature on clock skew could be modeled explicitly, which has a long history in the sensor array literature.

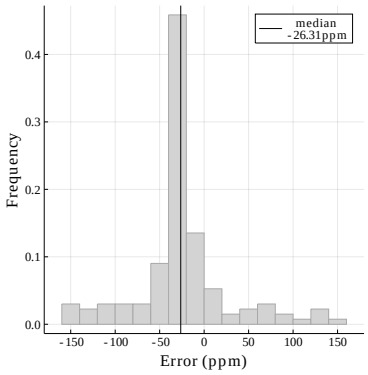

Figure 16: Histogram of clock skew estimates. Estimates outside ±150 were excluded.

[32] Nicholas J. Bryan, Miriam A. Kolar, and Jonathan S. Abel (2010). "Impulse Response Measurements in the Presence of Clock Drift"

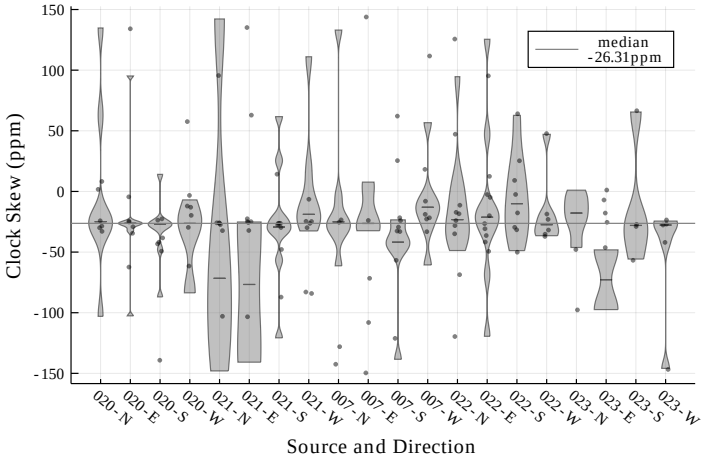

Figure 17: Summary of clock skew estimation for all impulse response recordings. Estimates outside ±150 were excluded.

Impulse Response Summary Statistics

To analyze the acoustic propagation at the field site, we developed a simple iterative algorithm to estimate the relevant statistics. Following the general model given by Traer and McDermott [33], we consider each frequency band of the impulse response separately, and assume that it contains an impulsive direct signal followed by an exponential decay, which is linear in the logarithmic domain.

[33] James Traer and Josh H. McDermott (Nov. 2016). "Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space"

Noise Floor The noise power is reported in dB without reference, but is comparable within the dataset.

Decay Time to -60dB (RT60) The time it would take for the reverberant tail to decay by 60dB. This is computed from the slope of the log-linear decay.

Direct-to-Reverberant Ratio (DRR) The ratio of the peak power to the reverberant power at the peak time, computed from by the log-linear decay.

Figure 18 shows an example result from this fit process, for a single band. In our analysis we implemented the filter bank via the STFT, with 1024-sample windows and 512-sample hop size. The fitting algorithm works by refining the boundaries of a region assumed to contain the decay (between the peak and the noise floor intercept - the time that the decay goes below the noise floor). The peak is determined by a simple maximum, and the noise floor intercept is initialized to be an over-estimate. A linear fit is performed in the decay region and the noise floor is estimated as the average power for the region after the intercept. A new noise floor intercept is determined by the intersection of the decay fit and the noise floor. The process is iterated until the noise floor intercept converges or increases.

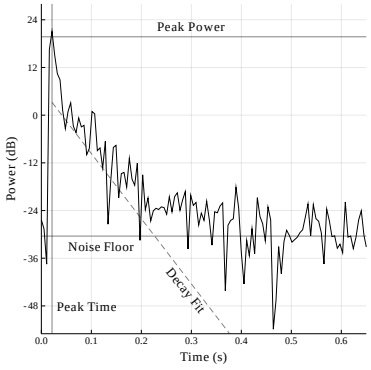

Figure 18: Example of impulse response fitting in a single band (here 1.78kHz).

We then ran the analysis on all our impulse responses, which included the MLS recordings as well as a number of direct impulses recorded with a slapstick. In total we collected 854 single-channel impulse responses. 221 have a peak power at least 30dB above the median power. Of these, 187 were recorded with the slapstick impulse and 34 were recorded with the MLS stimulus.

Figure 19 shows the results of the RT60 measurement, where we see the same general shape found in Traer and McDermott. Notice a substantial difference between the MLS and slapstick measurements. The MLS results align more closely with the prior work, which used a similar process, and showed RT60s generally below 0.5s for outdoor rural environments. It is not clear which measurement is more reliable, though there are fewer points of

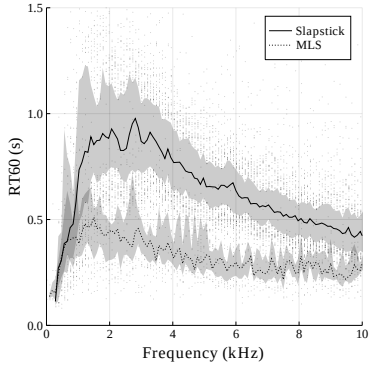

Figure 19: Plot of the RT60 for each band across the dataset. Recordings with an overall peak SNR below 30dB were removed because they did not provide accurate fits. Similarly frequencies above 10kHz often had very little impulse energy and so were ignored. The shaded area for each plot shows the 25th to 75th percentile range for each frequency band.

possible failure with the slapstick measurement. Two potential confounders are that the slapstick recordings were performed at different source locations on the site, and on a different day with less wind. In theory the RT60 should be independent of the SNR (assuming there is sufficient signal to accurately estimate it), but it is possible they are not sufficiently decoupled by the estimator. Another possibility is that the reverberation is due to time-varying factors that are removed by the averaging in the MLS process.

Recording Archive

We have accumulated over 10TB of audio data since 2012. Over 6TB is a multichannel stream stored in Ogg Opus format, with a total of 25,000 hours recorded (as of March 2020). During most of that time there have been between ten and fifteen microphones active. They are captured continuously to a 30-channel stream (the silent channels are compressed very efficiently by the Opus codec), as well as being streamed online as a stereo mix.

Because cables and other equipment have been damaged at times throughout the installation, not all of the recorded audio is usable. Water ingress often manifests as loud intermittent pops, clicks, and crackles, and open and short circuits frequently cause loud 60Hz hum. In the future we would like to train some simple classifiers to build a map of the data, so that it can be more easily incorporated into downstream applications, and also released as a dataset of more manageable size.

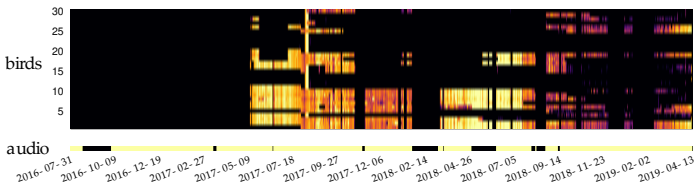

Figure 20: An incomplete map of data captured from the multichannel microphone array between July 2016 and April 2019.

Figure 20 shows an overview of the multichannel array data. The bottom line labeled "audio" highlights the regions where the system was capturing audio (black regions indicate outages). The "birds" heatmap indicates label density from a machine-learning classifier [34] that detects the presence of bird calls (as well as other labels). The classifier was occasionally modified throughout this monitoring period, so it's difficult to distinguish site variations from changes in the classifier. There does seem to be a visible drop in avian activity in the 2018/2019 winter, which is consistent with observations on site (though there are also two regions in that

[34] Clement Duhart et al. (Oct. 2019). "Deep Learning for Environmental Sensing Toward Social Wildlife Database". Paris, France

period where the classifier was not operating). Future work is planned to run a consistent classifier on the dataset as a whole.

Limitations and Future Work

Analyze IR Quality Additional work is to be done analyzing the impulse response dataset. One useful tool is to evaluate coherence, which measures the per-band correlation between input and output of a linear system. In this context the input is fixed and periodic, so coherence becomes simply the between-period variance of the response within each band. Listening tests performed by convolving these impulse responses with speech indicate that they are biased towards low frequencies, likely due to less-coherent high-frequency response, possibly due to errors in the skew estimation or actual time-variance in the system (both of which would disproportionally affect high frequencies).

Characterize Skew Estimation Further work is necessary to characterize the skew estimation procedure and compare more rigorously against prior work. Additional improvements are also likely available by pre-processing the data to reduce noise, and developing a heuristic to identify which channels are likely to provide useful skew estimates.

More survey data The data collected for this work used 22 different source locations and 13 different microphone locations, but a more systematic protocol would help resolve some remaining issues. The main issue is that the MLS recordings and slapstick recordings were performed on different days and different locations, so it is difficult to compare them. If it does turn out that time-variance plays a significant role in outdoor impulse responses, it would have significant impact on the methods used to capture them.

Part II

Acoustic Localization

Background: Cross-Correlation and Delay Estimation

The cross-correlation is a widely used building block for many DSP algorithms, though there are several subtleties that must be considered when using it in practice, and connections between disparate fields that can provide additional insight.

In signal processing, cross-correlation is often defined simply as convolution with one of the arguments time-reversed and conjugated. Equation 5 shows that the cross-correlation at time τ can also be framed as an inner product with one of the signals delayed by τ.

Convolution

$$(x * y)(t) = \sum_{\tau} x(\tau) y(t - \tau) \tag{4}$$

Cross-Correlation

$$(x \star y)(\tau) = \sum_{t} \overline{x(-t)} y(\tau - t) = \sum_{t} \overline{x(t - \tau)} y(t) \tag{5}$$

Here $\overline{x(t)}$ is the adjoint, or complex conjugate, of $x(t)$. We also change the variable to τ for cross-correlation because conventionally it refers to a time delay, not a moment in time. In this framework, the cross correlation takes two signals, giving a result that indicates their similarity (an inner product) for different time shifts, or *lags*. When cross-correlating a long signal with a shorter one, the peaks can be thought of as locating instances of the short signal within the longer one, a process often called *template matching*.

The Linear Algebra Perspective

Convolution and cross-correlation with a signal are linear functions and can be implemented by matrix multiplication.

We can define an operator $\mathcal{T}_n$ that takes a vector v and produces a matrix with n shifted copies of v as its columns, as in Equation 6.

$$\mathcal{T}_4\left(\begin{bmatrix}2\\1\\3\end{bmatrix}\right) = \begin{bmatrix}2&0&0&0\\1&2&0&0\\3&1&2&0\\0&3&1&2\\0&0&3&1\\0&0&0&3\end{bmatrix} \tag{6}$$

We can then express the convolution of vectors f and g, of lengths N and M respectively, as in Equation 7.

$$f * g = \mathcal{T}_M f g = \mathcal{T}_N g f \tag{7}$$

We can interpret this matrix multiplication as one vector giving the coefficients for a linear combination of shifted versions of the other (which is one way to describe convolution). Notice also that the number of rows in the convolution matrix (so also the length of the result) will be $L = N + M - 1$.

There are two main points of view when multiplying matrices, and the perspective depends on the problem being modeled. Consider the equation $y = Ax$. One way to think about this equation is that x gives a list of coefficients that define a linear combination of the columns of A. This is the perspective that leads to the interpretation of $\mathcal{T}_n$ above.

Rather than the column-focused interpretation, we can instead focus on the rows. Consider some matrix B, with $y = \overline{B}x$. Here again $\overline{x}$ denotes the adjoint, which is the complex conjugate for scalar values, and the Hermetian transpose for vectors and matrices. From this perspective, each element in the result gives the inner product between x and a column of B (for complex vectors the inner product is defined as in Equation 8).

$$\langle f, g \rangle = \overline{f} g = \sum_{i=1}^{N} \overline{f_i} g_i \tag{8}$$

Note that for a real-valued matrix-vector product Ax it's often convenient to think of the elements of the result as being the inner products of x and the rows of A, but for complex-valued A that interpretation is off by a conjugate, thus the slightly more complicated $\overline{B}$ formulation given here.

The cross-correlation $f \star g$ can be described as a collection of inner products between g and time-shifted versions of f. Combining the $\mathcal{T}$ operator introduced above with the inner-product interpretation of matrix products leads to the definition given in

Equation 9.

$$f \star g = \overline{\mathcal{T}_M f} g \tag{9}$$

This introduces an issue however, because $\mathcal{T}_M f$ is $L \times M$, so $\overline{\mathcal{T}_M f}$ cannot be multiplied by g directly. Notice however that we can add all-zero columns to $\mathcal{T}_M f$ while adding the same number of zero elements to g without changing the result. If we add these zero elements such that $\mathcal{T}_M f$ is square with dimensions $L \times L$ and g is length L, then Equations 7 and 9 both hold.

We can clean up some of this notational messiness by introducing a new operator $\mathcal{C}$ which is similar to $\mathcal{T}_n$ in that the columns of the result are shifted versions of the operand, but $\mathcal{C}$ uses a circular shift rather than adding additional rows, as seen in Equation 10.

$$\mathcal{C} \begin{bmatrix} 2 \\ 1 \\ 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 3 & 1 \\ 1 & 2 & 0 & 3 \\ 3 & 1 & 2 & 0 \\ 0 & 3 & 1 & 2 \end{bmatrix} \tag{10}$$

The result is known as a *circulant* matrix. With this, we can define convolution as $f * g = \mathcal{C} f g$ and cross-correlation as $f \star g = \overline{\mathcal{C} f} g$, under the assumption that f and g are both zero-padded to length L. If f and g are equal length but not zero padded, these same definitions give circular convolution and circular cross-correlation, respectively. To make the connection further, we can notice that the rows of $\mathcal{C} f$ are reversed and shifted versions of f, so multiplying by $\overline{\mathcal{C} f}$ is equivalent to convolving with time-reversed and conjugated f.

One interesting insight that arises from this formulation is that if the convolution matrix $\mathcal{C} f$ for some signal f is *unitary* (its adjoint is its inverse), then $\overline{\mathcal{C} f} \mathcal{C} f h = h$, implying that $f \star (f * h) = h$. If h represents the impulse response of some unknown system, we can excite the system with the stimulus f, giving $f * h$, and then use cross-correlation to extract h. This is reminiscent of the impulse response measurement techniques described in Part I. Another equivalent definition of a unitary matrix is one where the columns form an orthonormal basis for $\mathbb{C}^L$. Recall that the columns are shifted versions of f, so f is orthogonal to all shifted versions of itself (i.e. f's autocorrelation function is an impulse). This is a restatement of the "shift-orthogonality" property that motivates the use of Maximum Length Sequences (which are approximately shift-orthogonal, and the Random Phase Multisine method.

The Statistics Perspective

In probability and statistics it's common to think about *correlation* and *covariance*, which describe how multiple random variables vary together[35]. The correlation of two random variables is defined as $\mathbb{E}\left[Y\overline{X}\right]$. This definition applies whether X and Y are scalars or vectors. For random vectors X and Y, this gives an outer product, and the result is a matrix R_{XY}, with $R_{XY}(i,j) = \mathbb{E}\left[Y_i\overline{X_j}\right]$.

Similar definitions exist for *covariance*, which require subtracting the mean from the signals of interest. In a signal processing context we are often dealing with zero-mean signals, in which case the covariance and correlation are equivalent. Also note that some authors define correlation as a normalized covariance. Here however we use the terminology of Oppenheim and Verghese [36], which refers to the normalized covariance as the *correlation coefficient*.

Complex-valued random variables also require an additional metric called the *pseudo-covariance* [37], defined as $\mathbb{E}\left[YX^T\right]$ (note the non-hermetian transpose) to be fully-specified. However, this can be (and often is implicitly) ignored for random variables that are *circularly-symmetric*, i.e. their real and imaginary parts are uncorrelated.

The terms *cross-correlation matrix* and *autocorrelation matrix* (and sometimes without the "matrix" qualifier) are frequently-used when the random variables in question are signals, though the above definition of correlation still applies. Autocorrelation refers to the case where $X = Y$ (which typically isn't named in the scalar case). In this case the correlation matrix is square, hermetian, and has the variance of each component along the diagonal.

Wide-sense stationary (WSS) signals (and pairs of jointly-WSS signals) permit a further simplification: because the correlation $R_{XY}(i,j)$ depends only on the difference $(i-j)$, it is often written as a function of a single variable as in Equation 11.

$$R_{XY}(\tau) = R_{XY}(i, i-\tau) = \mathbb{E}\left[Y_i\overline{X_{i-\tau}}\right] \tag{11}$$

This is equal for all i, which implies that for jointly WSS signals the correlation matrix R_{XY} has Toeplitz structure (constant diagonals).

Recall that the signal processing definition of the cross-correlation is $(x \star y)(\tau) = \sum_t \overline{x(t-\tau)}y(t)$. This can be interpreted as an empirical estimate of the expectation in Equation 11 (modulo a scaling factor). So the DSP cross-correlation estimates the diagonals of the correlation matrix.

This perspective also informs normalization. If x and y are length N and M signals, respectively, that are assumed to come

[35] This description has an important caveat: covariance and correlation only describe linear relationships between variables. For example, if X is a random variable and $Y = X^2$, their correlation is zero though they are deterministically related. They are *uncorrelated* yet not independent.

[36] Alan V Oppenheim and George C Verghese (2010). *Signals, Systems, and Inference: Class Notes for 6.011: Introduction to Communication, Control and Signal Processing Spring 2010*

[37] Robert G Gallager (2008). *Circularly-Symmetric Gaussian Random Vectors*. Tech. rep.

from some jointly WSS random process, the scaling factor needed to estimate the correlation matrix from the cross-correlation varies with the lag, based on the amount of overlap between the two signals. Consider the case where $M = N$. At zero-lag, the signals exactly overlap so $\widehat{R}_{XY}(0) = \frac{1}{N}(x \star y)(0)$. For nonzero lag only a portion of x and y overlap, giving the more general expression in Equation 12.

$$\widehat{R}_{XY}(\tau) = \frac{1}{N - |\tau|}(x \star y)(\tau) \tag{12}$$

This equation is valid for $-N < \tau < N$. Notice that as $|\tau|$ approaches N there are fewer observations being averaged, increasing the variance of the estimator. In this context it's often best to ignore the edges of the correlation as they are not reliable and can be high-magnitude relative to the rest of the signal.

This is somewhat related to "Time-Varied Gain" as used in sonar rangefinding, except that in that context the energy loss is due to propagation, and in this case it is due to the reduced signal overlap. Note that this scaling factor is only appropriate for signals where the correlation is largely time-invariant within the analysis frame. In a delay estimation context this corresponds to signals where the target occupies most of the frame. If the target signal is short-duration, such that it is entirely within the overlap of both signals for the relevant lag range, this correction is not appropriate and will bias against zero-lag.

Time Aliasing

Given signals x and y, the cross-correlation can be computed efficiently with the Fourier transform:

$$(x \star y) = \mathcal{F}^{-1}(\overline{\mathcal{D}(\mathcal{F}x)}(\mathcal{F}y)) \tag{13}$$

Where $\mathcal{D}\cdot$ creates a diagonal matrix, so $\overline{\mathcal{D}(\mathcal{F}x)}(\mathcal{F}y)$ is an elementwise product.

In a DSP context $\mathcal{F}$ represents the Discrete Fourier Transform (DFT), where the signal is considered to be periodic in both time and frequency. As given, this operation will perform a circular cross-correlation - as one signal is shifted relative to the other it will wrap around. Often this is undesirable (e.g. if the signals are not actually periodic), and zero-padding is required to avoid *time-aliasing*.

The Generalized Cross-Correlation

There is a rich body of literature on improved cross-correlation variants, generally with the goal of sharpening the cross-correlation peak so that the time difference is easier to estimate. The Generalized Cross Correlation adds a filter to the cross-correlation process, and has been previously used in gunshot detection systems [38]. There has also been recent work [39] that estimates the same time-frequency mask used for source separation as a pre-processing step to improve cross-correlation.

[38] Giuseppe Valenzise et al. (2007). "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems"

[39] Zhong-Qiu Wang, Xueliang Zhang, and DeLiang Wang (Sept. 2018). "Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks"

The basic assumption behind delay estimation using cross-correlations is that the delay between channels will appear as a peak in the cross-correlation function. Finding the peak can be made more challenging in the presence of noise, or because of oscillations caused by strong narrowband components. The Generalized Cross-Correlation (GCC) [40] adds a pre-filtering step with the goal of making the peak more easily detectable. This is generally expressed in the frequency domain as:

[40] C. Knapp and G. Carter (Aug. 1976). "The Generalized Correlation Method for Estimation of Time Delay"

$$(x \star_\Phi y)(l) = \int \Phi(\omega)\overline{\widetilde{x}(\omega)}\widetilde{y}(\omega)e^{j\omega t}d\omega \tag{14}$$

Where Φ is the filter, which is generally a function of x and y. With the Phase Transform (GCC-PHAT) approach, pre-filter is designed such that the magnitude spectrum of the cross power spectral density (the Fourier transform of the cross-correlation) is set to 1.

$$\Phi_{x,y}(\omega) = \frac{1}{|\overline{\widetilde{x}(\omega)}\widetilde{y}(\omega)|} \tag{15}$$

One way to understand GCC-PHAT is that it's a cross-correlation where we weight all the frequencies equally, so peaks in the cross-correlation occur where the most frequency bands are consistent with that lag. Often a signal has strong periodic components that create oscillations in the cross-correlation, making it more difficult to identify the peak. Frequencies with low energy are not necessarily less valuable for delay estimation (as long as they are not corrupted by noise). It's common for real signals to be dominated by their low-frequency components, which tends to spread out the cross-correlation peak, so applying GCC-PHAT is widely used to sharpen the peak. In this context one can think of the regular cross-correlation as weighting each frequency's importance to the cross-correlation by its energy.

Figure 21 shows the cross-correlation compared against GCC-PHAT. At this time scale it does not appear that GCC-PHAT offers much of an improvement (both correlations have sharp peaks).

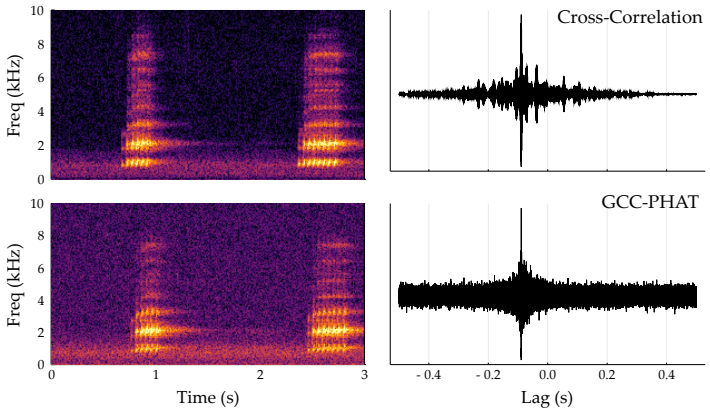

Figure 21: An example of the regular Cross-Correlation compared against GCC-PHAT. The spectrograms at the left show signals from two different microphones.

When we look more closely however, as in Figure 22, we can see that the regular cross-correlation has strong oscillations that are not present in the GCC-PHAT. Counting the oscillation cycles, we see that they correspond to a periodic component at around $2kHz$, which corresponds to the dominant frequency visible in the spectrograms.

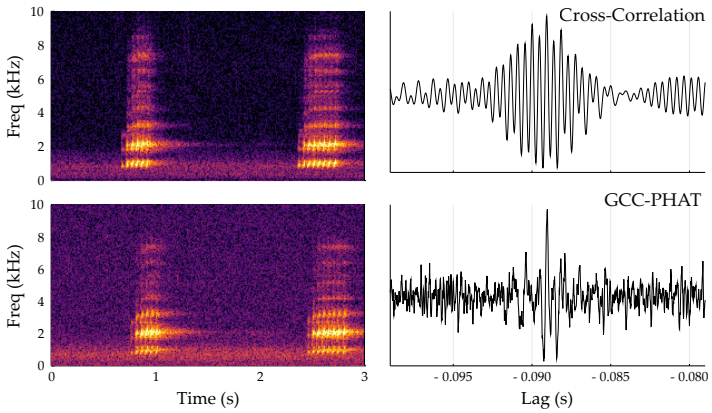

Figure 22: The same signals as Figure 21, but zoomed into a much smaller time range. .

Background: TDoA and Nearfield Localization

Acoustic localization (determining the location of an audio source) is important in two distinct ways for this project. Most directly, the location estimate determines where to place the virtual sound source in the auditory scene. Additionally, the location of the sound informs the source separation process by providing a geometrically-consistent estimate of the time differences between channels. This can be used to time-align the channels, including the channels without significant signal energy.

The previous chapter focused on delay estimation between pairs of microphones. This chapter describes using those delays to estimate the location of a source using a Time Difference of Arrival (TDoA) method. One approach would be to perform pairwise delay estimation between all pairs of microphones (for example, using the peak of the cross-correlation function). The location is then estimated as the one that is most consistent with the delays. This is referred to as an *indirect* approach [41]. The downside to this is that it forces a choice of delay for each pair - for instance, if an interfering source causes dominant peaks in several microphone pairs, any secondary peaks due to the target source are ignored completely.

[41] M. Cobos, A. Marti, and J. J. Lopez (Jan. 2011). "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling"

Rather than rely on peak picking for each pair of input channels, *direct* methods integrate information from all microphone simultaneously. One benefit to the direct approach is that it does not require *a priori* knowledge of which channels contain the target signal. This information is often used to pick a particular channel as a reference channel [42], or to avoid introducing noise from noisy channels. In a large distributed array it is likely that only some subset of the microphones have significant target energy relative to the noise, but we don't know which ones they are. Because of this, it is helpful to be able to integrate all the signals without needing a channel selection step.

[42] Keisuke Hasegawa et al. (2010). "Blind Estimation of Locations and Time Offsets for Distributed Recording Devices". Ed. by Vincent Vigneron et al. Lecture Notes in Computer Science. Berlin, Heidelberg

To better understand these techniques, first consider a single pair of microphones at positions m_1 and m_2. For a source at position p, the time difference of arrival (the delay between when

the source signal arrives at m_1 and m_2) is given by $\Delta t_{1,2}(p) = \frac{1}{c}\left(\|p - m_2\| - \|p - m_1\|\right)$, where c is the speed of sound. Note that this applies whether in 2D or 3D.

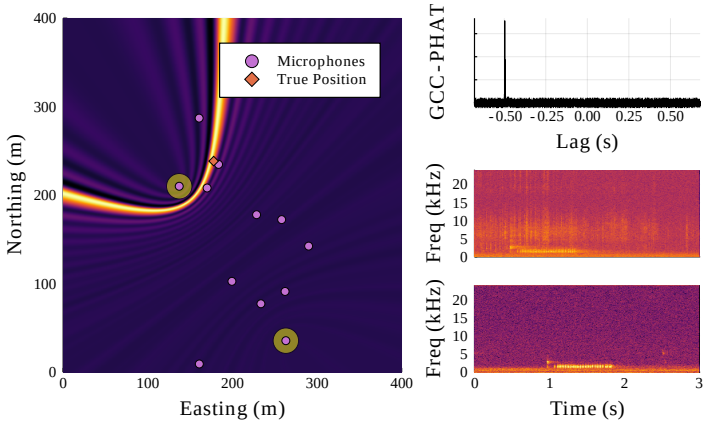

Figure 23: This plot shows a heatmap of the proposed Spatial Likelihood Function, using only the highlighted microphone pair. The heatmap shows evaluations of the function on a $1m$ grid in a 400×400 m^2 neighborhood of the microphone installation. The source audio is a bird call played from a speaker at a known location, recorded by the array. The spectrograms on the lower-right display the channels highlighted in the heatmap. On the upper right the squared cross-correlation is displayed, computed using the GCC-PHAT algorithm.

The TDOA $\Delta t_{1,2}(p)$, is not unique to p, but is shared by all points on a hyperboloid that passes through p and is symmetric around the line through m_1 and m_2. This is visible as the highlighted hyperbola in Figure 23. Two special cases are worth highlighting: if $\Delta t_{1,2}(p) = 0$, then the signal arrived at the microphones simultaneously, indicating that p must be somewhere on the plane bisecting m_1 and m_2. This can be seen in Figure 26, where we observe a peak at zero-lag in the cross-correlation, as well as a straight line highlighted, bisecting the microphones. Note that in this case the zero-lag source is not the dominant one. If $\Delta t_{1,2}(p) = \pm\frac{\|m_2 - m_1\|}{c}$ (i.e. the largest possible delay), then p must lie on a beam emanating from m_1 or m_2, going directly away from the other microphone. This condition is visible in Figures 25 and 26. This is the degenerate case of the hyperboloid where the bend becomes infinitely sharp and it becomes a beam.

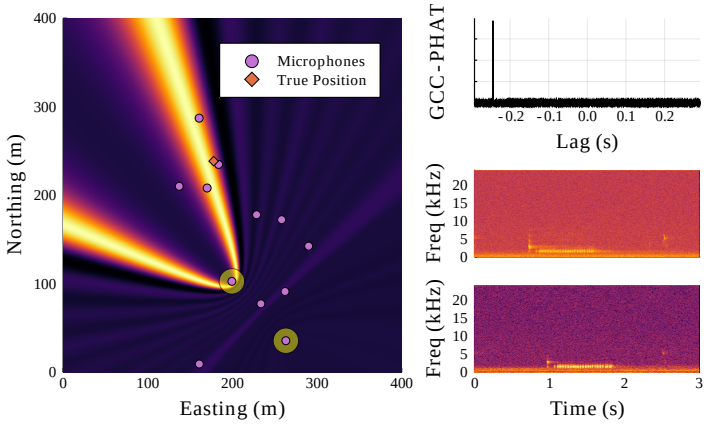

Figure 24: The spatial likelihood function, as in Figure 23, but with a different pair of microphones.

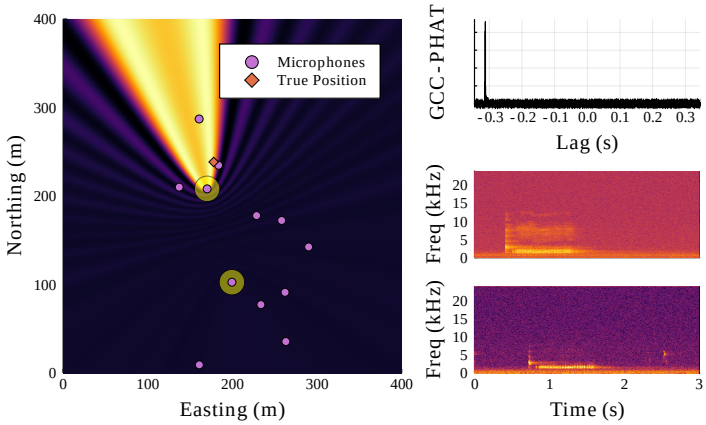

Figure 25: The spatial likelihood function, as in Figure 23, but with a different pair of microphones.

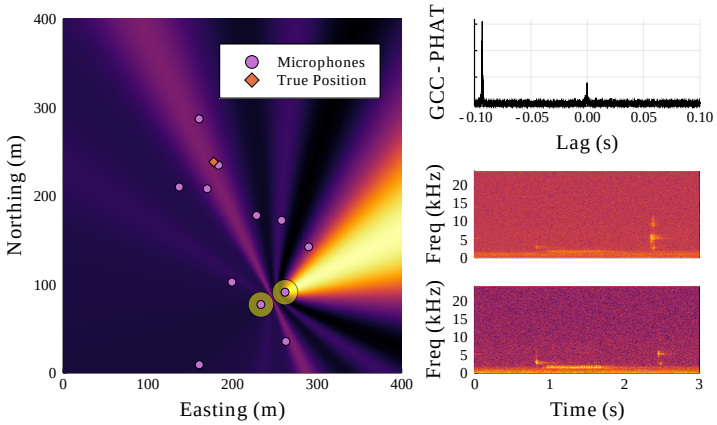

Figure 26: The spatial likelihood function, as in Figure 23, but with a different pair of microphones.

Steered Response Power

One important direct method is known as the *steered response power* (SRP). The SRP is a function of location (or direction in a farfield direction of arrival context) that computes the output power for a given multichannel input with the array is focused at that point. To focus the array, the microphone signals are delayed appropriately based on a known speed of sound and distance from the source position. The delayed signals are then summed, and the power of the resulting mixture is computed. If the hypothesized location corresponds to the source's actual location, and the transfer functions from the source to the microphones differ only by a pure delay, the delayed microphone signals should sum coherently, so the SRP should be maximized at the source location.

We can define the SRP function for microphones with signals $\{x_1, x_2, ..., x_I\}$:

$$\mathrm{SRP}(p) = \frac{1}{N} \sum_{n=1}^{N} \left| \sum_{i=1}^{I} x_i\big(n - \Delta t_{1,i}(p)\big) \right|^2 \tag{16}$$

The first microphone is arbitrarily chosen to be the reference that all other signals are shifted against. Note that the shifts are based on the geometry of the array and the hypothetical position, not the content of the signals. When computing this function at a large number of points, it becomes more efficient to use an alternate formulation that is given in terms of the generalized cross-correlations of all the microphone pairs[43]:

$$\mathrm{SRP}'(p) = \sum_{i=1}^{I} \sum_{j=i+1}^{I} (x_i \star x_j)\big(\Delta t_{i,j}(p)\big) \tag{17}$$

This formulation isn't completely equivalent, but captures the portion of $\mathrm{SRP}(p)$ that varies with p. With this method the cross-correlations can be pre-computed, and evaluating the function at each point is $\frac{N(N-1)}{2}$ table look-ups (one for each cross-correlation) summed together.

One major issue with the SRP framework is that it assumes that the transfer functions between the source and microphones differ only by a delay, so after compensating for the delay the signals add constructively. When the microphones are far apart this property does not hold. To avoid these phase errors, I experimented with a variant of SRP where the term under the summation was the power of the cross-correlation, but found that this required a variety of ad-hoc weighting factors and normalization for acceptable performance.

[43] Joseph Hector DiBiase (May 2000). "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays". PhD. Providence, RI

Spatial Likelihood Function

Alternatively, we can define a *spatial likelihood function* (SLF) [44]. The SLF treats the source location as a model parameter, for which we can compute a likelihood given the observed microphone signals. In practice the models used for SLFs often generate very similar formulations to the SRP, but the probabilistic framework provides additional flexibility. Aarabi also includes a "spatial observability function" that accounts for the distance between each microphone and the source when computing the SLF. However, even distant microphones can often contribute useful information, and the observability for a given microphone would need to be a function of the local noise conditions at the microphone and the spectrum of the source.

[44] Parham Aarabi (2003). "The Fusion of Distributed Microphone Arrays for Sound Localization"

Proposed Spatial Likelihood Function

The SLF is quite general, and allows a variety of probabilistic models to be used. As noted by Aarabi, we're generally interested in comparisons in likelihood between different locations, so it suffices to estimate some monotonically function of the likelihood, rather than the likelihood itself. In this work we consider the simple model where the GCC-PHAT cross-correlation between each microphone pair represents an independent observation. The noise variance of the cross-correlation is estimated as the median power, so it is robust to cross-correlation peaks. We then compute a likelihood for each sample

$$\mathrm{SLF}(p) = \prod_{i=1}^{I} \prod_{j=i+1}^{I} f_{\mathcal{N}} \left(x_i \star_{\Phi} x_j \left(\Delta t_{i,j}(p) \right) | \sigma_{i,j} \right)^{-1} \tag{18}$$

Where $f_{\mathcal{N}}(\cdot|\sigma)$ is the PDF of a normal distribution with standard deviation σ, and $\sigma_{i,j}$ is the standard deviation of the cross-correlation between channels i and j. Because the PDF gives the likelihood under a noise model, we take the inverse so the estimated source location corresponds to the maximum of this SLF, rather than the minimum. In practice we compute this in the logarithmic domain, in which case it becomes a nested summation that looks very similar to Equation 17. This seems to work sufficiently well in practice and did not require ad-hoc weighting schemes, though performance could likely be improved with a more sophisticated model that accounts for the peak location explicitly. This would address one of the current weaknesses which is that high-amplitude cross-correlation peaks can sometimes dominate the response. Figure 27 shows the proposed spatial likelihood function using all microphones.

Handling Aliasing

The SLF is evaluated on a 2D grid approximately on the plane of the microphones. The result is a heatmap showing where the

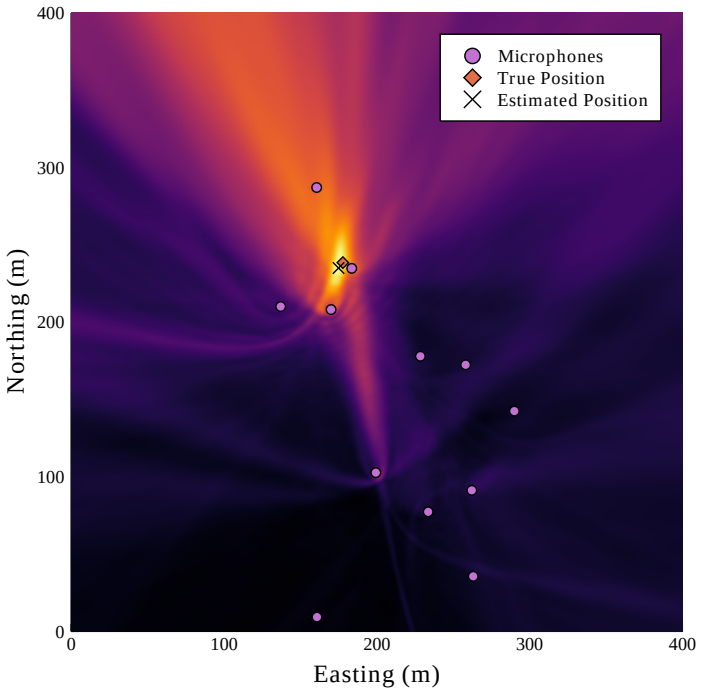

Figure 27: A heatmap generated by sampling the SLF function for audio recorded in the field at Tidmarsh, using all pairs of microphones. The ground truth location is plotted, as well as the estimated position (the maximum of the SLF), and the microphone locations. Dimensions are in meters.

source is likely to be under the model. This can be interpreted as a 2D sampling of the continuous SLF function, which has much greater bandwidth than is reasonable to sample. With a grid spacing of 1m, the distance diagonally across the grid is $\sqrt{2}$, which corresponds to sampling the cross-correlation at 225Hz, which would cause severe aliasing if not addressed. Intuitively the issue is that the cross-correlation peaks are usually very thin when visualized on the 2D heatmap, and can be missed by the sampling grid. To address this we lowpass filter and downsample the cross-correlation power for each pair before computing the Gaussian PDF (the PDF is then evaluated on the square root of the smoothed power).

Ground Truth Data Capture

To test the accuracy of the proposed localization system, as well as the quality of source separation, we collected ground truth data - known signals played from known locations. The signals consisted of 12 recordings spanning a variety of signal types, including bird, amphibian, insect, and human vocalizations. They were also selected to span a range of broadband and narrowband characteristics, with some being percussive in nature and others harmonic. As with the impulse response recordings, these were played through a Bose S1 portable speaker, and each playback location was measured using differential GPS.

Bird recordings were from the Cornell Guide to Bird Sounds: Master Set for North America [45]:

- American Crow - 143215301
- Blue Jay - 43214771
- Canada Goose - 43183481
- Chipping Sparrow - 43233791
- European Starling - 43224221
- Red-winged Blackbird - 43240421
- Semipalmated Sandpiper - 43196781

Speech sounds came from the LJ Speech Dataset [46]:

- Speech LJ037-0171
- Speech LJ025-0076

We also used several sounds from Freesound.org:

- cricket [47]
- tree frog [48]
- frog, lakeside [49]

Each file was edited to roughly 8 seconds, and in some cases cleaned up using RX7 Spectral Repair from iZotope, Inc. to remove background noise. They were also each normalized to the BS.1770-2/3/4 Loudness standard, to an integrated loudness of -24 LUFS with true peaks limited to odBFS (full scale), again with RX7.

[45] *Cornell Guide to Bird Sounds: Master Set for North America* (2014). Ithaca, New York

[46] Keith Ito (2017). *The LJ Speech Dataset*

[47] FunkApache (2017). *393389__funka-pache__cricket.wav*

[48] alienistcog (2014). *241974__alienistcog__2014-tree-frogs3.aiff*

[49] kayceemixer (2014). *251495__kayceemixer__kc-animal-frog-lakeside-penticton-2013.wav*

Results

To characterize the performance of the localization algorithm, we measure the error between the estimated location and the ground truth location for each recording in the naturalistic audio dataset. The ground truth was measured with the RTK GPS and is accurate to within approximately 50cm. The likelihood function was evaluated with $1m$ resolution within the surveyed area. For each recording we also computed a quality metric, sorting all the channels by their source-to-distortion ratio (SDR) and taking the mean of the first three. This is because 2D localization requires the signal to be present in at least 3 microphones. We refer to this metric as SDR_3. SDR for each channel was computed using the `mir_eval` python package [50], which measures the amount of energy in the mixture that can be explained by the target, allowing for convolution by a 512-point FIR filter. That is, it projects the mixture into the subspace spanned by 512 time-shifted versions of the target signal, and gives the ratio of the energy in the subspace to the energy in the orthogonal subspace.

[50] Colin Raffel et al. (2014). "MIR_EVAL: A Transparent Implementation of Common MIR Metrics". Ed. by Hsin-Min Wang, Yi-Hsuan Yang, and Jin Ha Lee

Figure 28 shows that the localization is mostly random below -12dB SDR_3, and by -6dB it has become quite accurate. Figure 29 shows only the lower-right portion of the results plot, and shows that the error converges to about four meters. This remaining error is likely due to a combination of measurement errors in the microphone positions, errors due to the 2D assumption (in reality there are a few meters elevation difference between microphones), and quantization error in the spatial likelihood evaluation.

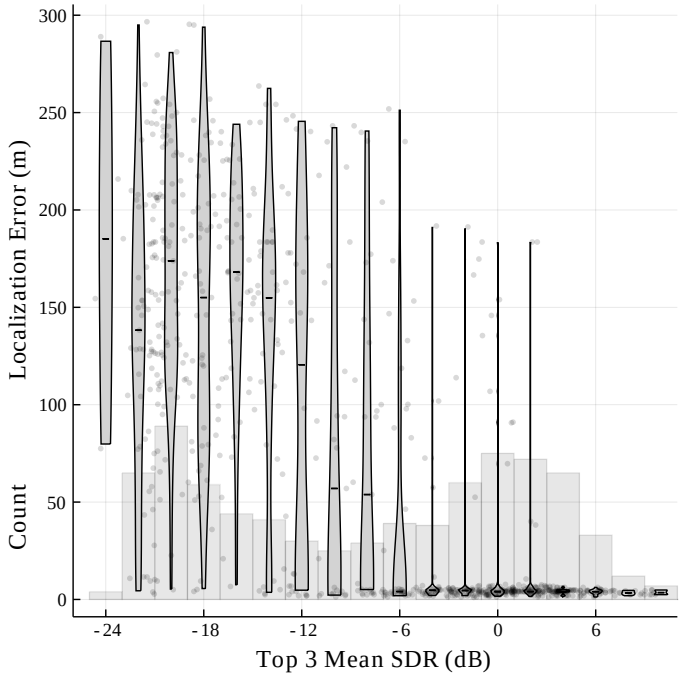

Figure 28: Localization results. Each data point is a multichannel recording from the naturalistic recordings dataset. Location accuracy is plotted against the mean SDR of the best 3 channels, under the assumption that the quality in several channels is important for high-quality cross-correlations and localization. Violin plots show the data quanitized to 2dB increments, to better visualize trends, and light grey bars show a histogram with the total number of results in each bin.

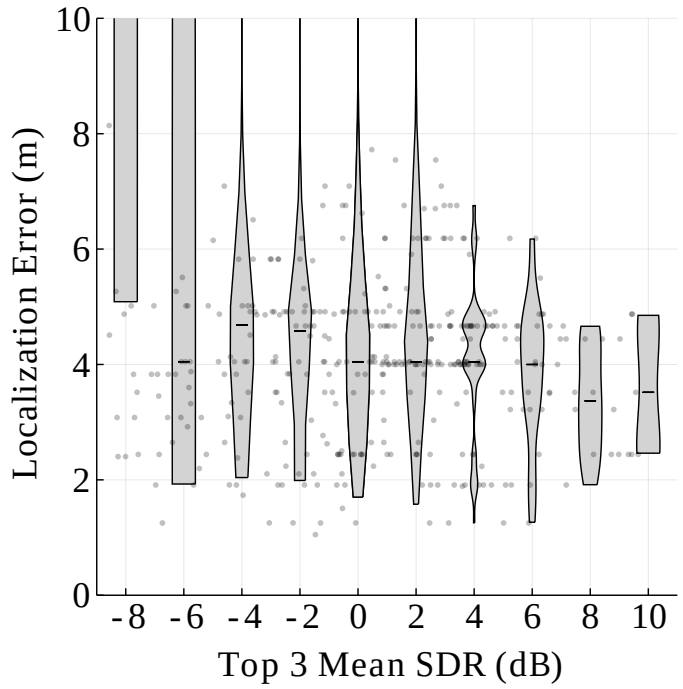

Figure 29: The same data as Figure 28, but focused on the area where localization was successful.

Limitations and Future Work

Elevation The primary limitation in this implementation is that it does not estimate source elevation, and in fact elevated sources will degrade the system's ability to estimate their position projected onto the ground plane. Considering a volume rather than a surface of potential source locations would require a different approach because the state space would be too large to discretize and exhaustively check, so an iterative optimization scheme would likely be required.

Array Shape Calibration The systems precision is limited by the precision of the microphone location measurements, which can be improved through array shape calibration. This is a well-studied problem, and solutions exist to calibrate sensor locations even using unknown sources (providing there is redundancy in the array and sufficient strong sources) [51].

There is a limit to the precision improvements possible through calibration however. Further improvements may require extra parameters such as wind speed and direction, and temperature.

Improving Noise Robustness These localization results provide a target (roughly -6dB) for what the SDR needs to be in order for acceptable localization. The system could be made more robust to noise by pre-processing the individual channels prior to the cross-correlation to bring more channels above this threshold. This pre-processing would need to incorporate assumptions about the target signals, and would be a good candidate for machine learning, where single-channel source separation techniques continue to improve.

Specifically, GCC-PHAT helps reduce the influence of periodic components in a cross-correlation, but also amplifies noise. With an estimate of the signal and noise power spectra, a Wiener-like scaling term could be added to the PHAT pre-filter, so that low-SNR frequencies get removed rather than amplified. That is, we

[51] Y. Rockah and P. Schultheiss (June 1987). "Array Shape Calibration Using Sources in Unknown Locations–Part II: Near-Field Sources and Estimator Implementation"

could modify the standard GCC-PHAT filter as follows:

$$\Phi_{wph}(\omega) = \frac{P_s(\omega)}{P_s(\omega)|X(\omega)| + P_n(\omega)} \tag{19}$$

Here P_s and P_n are signal and noise power estimates, respectively, and X is the cross-correlation spectrum.

Localization Probability Model Currently the probabilistic model for the spatial likelihood model is very rudimentary (just using the noise estimate and looking for *unlikely* locations). This still struggles to make the best use of low-amplitude peaks, which could be just as valuable as the high-energy peaks, given a more sophisticated model that models the peaks explicitly. Improvements in the model would likely drive the SDR threshold for successful localization downwards, as the model would make better use of low-SDR data.

Moving Sources The localization system currently assumes that sources are stationary within a given analysis window (3 seconds in these experiments). In this implementation, moving sources would likely manifest as broader peaks in the cross-correlation (and subsequently the spatial likelihood function). One simple approach would be to use shorter analysis windows to reduce the period over which the sources are assumed stationary. However, there is a trade-off because longer windows provide more context and better noise performance. The maximum inter-channel delay is roughly 1 second, so the analysis window must be longer than that to ensure that a given source is present in all microphones. This limitation could be addressed by using different signal windows for different regions of the SLF. This would permit shorter cross-correlations, but more of them would be necessary because they wouldn't be shared across the whole SLF.

Part III

Foreground/Background Separation

Background: Subspaces and Matrix Factorization

Often a signal of interest can be defined in terms of a linear combination of a small number of component signals. A length-N signal has N degrees of freedom, but if it can be described as the sum of M components, where $M < N$, it has a more compact description with only M degrees of freedom. For example, a real-valued sinusoidal signal with known frequency but arbitrary phase and amplitude can be represented as a linear combination of sin and cos terms. In other words, all such signals can be thought of as vectors that lie in a 2D subspace, with $\sin(\omega t)$ and $\cos(\omega t)$ as basis functions. With this model the signal can be described with just two degrees of freedom. If such a signal is observed mixed with noise, the mixture can be thought of as a linear combination of a component within the signal subspace, and an orthogonal component due to the noise. Projecting the mixture into the signal subspace can thus be considered a denoising operation. This signal model can be represented as $x = s + n = Av + n$, or

$$\begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(N-1) \end{bmatrix} = \begin{bmatrix} A_0(0) & A_1(0) \\ A_0(1) & A_1(1) \\ \vdots & \vdots \\ A_0(N-1) & A_1(N-1) \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} + \begin{bmatrix} n(0) \\ n(1) \\ \vdots \\ n(N-1) \end{bmatrix} \tag{20}$$

Where the columns of A are $\sin(\omega t)$ and $\cos(\omega t)$. For a general A, the matrix that projects into the signal subspace (the column space of A) is given by $P = A(\overline{A}A)^{-1}\overline{A}$ [52]. If the columns of A are orthogonal to each other (as they are in this example), then $\overline{A}A$ is a diagonal matrix with the energy of each basis function on the diagonals, so $(\overline{A}A)^{-1}$ simply gives a normalizing factor for each component. If the columns are also normalized (which they are not in this example) than $\overline{A}A$ is the identity matrix and the projection simplifies to $P = A\overline{A}$.

Figure 30 shows an example of this process. It should not be surprising that it is possible to recover a good estimate of the

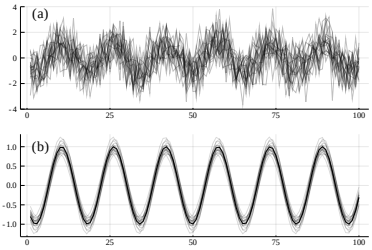

Figure 30: Subspace denoising of a sinsoid at a known frequency in white gaussian noise, with 20 random noisy mixtures. (a) shows the noisy mixture and (b) shows the signal estimates (grey) and true signal (black).

[52] Note that if A is not invertible then A^{-1} does not exist, so we can not simplify with $(\overline{A}A)^{-1} = A^{-1}\overline{A}^{-1}$. If A is invertible, then its columns span the whole space so P is the identity matrix.

original signal through substantial noise, because we are using a tightly constrained model of the signal (that it is a sinusoid at a given frequency). However, when an accurate model is available, and when that model can be represented as a linear combination of basis functions, projection is a powerful but straightforward way to take advantage of the model.

The applicability of this technique would be limited if one always required the signal subspace *a priori*. Fortunately a trade-off is available: if multiple (preferably many) instances of the noisy mixture are available, the basis can be estimated from the data, as seen in the next section.

Principle Component Analysis

The structure of the correlation matrix can provide valuable insight on the distribution of the data. First consider that for a particular deterministic vector x, $x\overline{x}$ gives a matrix where the ith column is $x\overline{x_i}$. Because each column is a scaled version of x, it is rank-1. The correlation is an expected value $\mathbb{E}\left[X\overline{X}\right]$, so it is a weighted average of all such rank-1 matrices. If all observations of a length-N vector X are multiples of the same vector x_0, then the correlation matrix will itself be rank-1. In general if all observations of X are linear combinations of M vectors for $M \leq N$, the correlation matrix will be rank-M.

As an example, consider a hypothetical vector-valued random variable X of uncorrelated data, as seen in Figure 31. The correlation matrix $\mathbb{E}\left[X\overline{X}\right]$ is a diagonal matrix giving the variance of each component.

Now assume we're unable to observe X directly, but only a related variable $Y = AX$ (where Y is significantly higher-dimensional than X). So each observation of Y is a linear combination of the columns of A, with the coefficients given by X. Notice that the columns of A corresponding to the high-variance components of X will contribute more to the distribution of Y. In the limiting case where only one component of X had nonzero variance, each sample from Y would just be a scaled version of the corresponding column of A. This property where the observed quantity is a linear combination of a small number of (often unknown) components is known as *low-rank structure* - one primary application of PCA is to recover this structure from the observed data. Another way to think about this property is that because the observed data is made of linear combinations of the columns of A, the data lies in a lower-dimensional subspace spanned by those column vectors, and we should be able to identify that subspace.

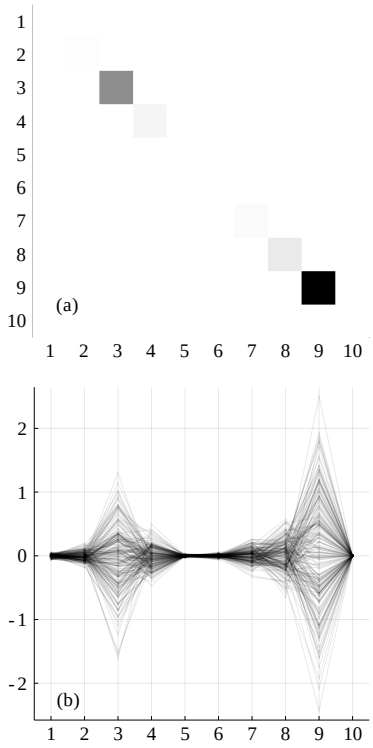

Figure 31: Observations of a 10-dimensional gaussian distribution with diagonal correlation. (a) gives the correlation matrix (darker colors are greater correlation) and (b) displays 200 observations as waveforms.

Figure 32 shows an example derived from X with A given by a random 100×10 matrix.

The correlation matrix of Y is given by

$$\begin{aligned} R_{YY} &= \mathbb{E}\big[Y\overline{Y}\big] = \mathbb{E}\big[AX\overline{AX}\big] \\ &= \mathbb{E}\big[AX\overline{X}\,\overline{A}\big] \\ &= A\,\mathbb{E}\big[X\overline{X}\big]\,\overline{A} \\ &= A\,R_{XX}\,\overline{A} \end{aligned} \tag{21}$$

So the question is - Given R_{YY} (which we can estimate from our observations of Y), what can we learn about A and X?. We can perform an eigenvalue decomposition on R_{YY} to get

$$R_{YY} = Q\Lambda Q^{-1} = Q\Lambda\overline{Q} \tag{22}$$

Where Λ is a diagonal matrix giving the eigenvalues (which are non-negative and real-valued), and the columns of Q give the eigenvectors of R_{YY}. $Q^{-1} = \overline{Q}$ (Q is unitary) because R_{YY} is hermetian. Thus Y is equivalent to a distribution with diagonal correlation Λ, linearly transformed by the unitary matrix Q.

Recall that the goal was to recover the low-rank structure that was used to generate Y. However, while the structure in Equations 21 and 22 are similar, in general $Q \neq A$ and $\Lambda \neq R_{XX}$. One difference is that the ordering of the eigenvectors and their corresponding eigenvalues could be permuted. More importantly, the columns of Q are not the same as the columns of A.

While the eigendecomposition cannot recover A and x independently, it *does* provide an orthogonal basis for the column space of A, which is useful for projecting into that subspace. The eigenvalues indicate how much variance is explained by each eigenvector, which can be used to estimate the rank of the subspace containing the signal, and to choose which components are most important to include when performing a low-rank approximation.

The eigenvectors are in fact the *principle components*, and in practice are typically sorted in descending order of their eigenvalues. In low-dimensional spaces these are often thought of as the directions of most variance, but in a signal processing context it makes sense to shift perspective and think instead of the principle components as signals whose linear combinations form the observed data.

Figure 33 shows the eigenvalues from Λ. Compare to the variances visible in 31, though as mentioned, the order is lost.

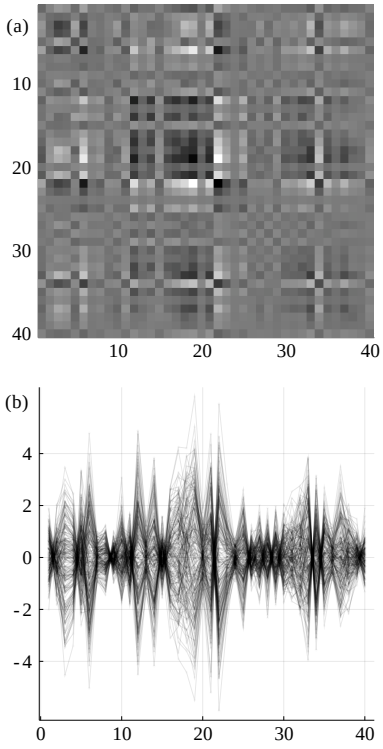

Figure 32: 200 samples of a 100-dimensional gaussian distribution constructed as a linear transformation of the data plotted in Figure 31. Note the plaid pattern typical of low-rank matrices.

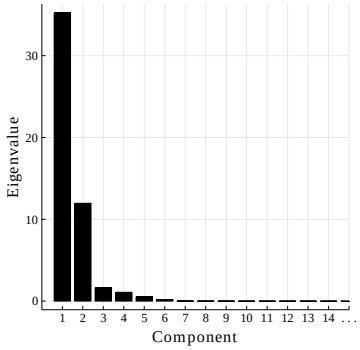

Figure 33: Eigenvalues of the covariance matrix R_{YY}.

Singular Value Decomposition and Low-Rank Approximation

In the previous section we described Principle Component Analysis, and showed how the principle components can be computed from the eigendecomposition of the covariance matrix. In practice the covariance matrix is not generally given, but instead is estimated from data. Here we define D as a $N \times M$ matrix with M observations of an N-dimensional random variable X. The empirical estimate of the $N \times N$ correlation matrix is then given by $\widehat{R}_{XX} = \frac{1}{M} D\overline{D} = Q\Lambda\overline{Q}$ (again using the eigendecomposition). Recall that if the observations (columns of D) are linear combinations of L components, they live in a subspace spanned by the first L principle components. Define Q' as the matrix of the first L columns of Q. Because Q is unitary, its columns (and those of Q') are unit-norm and orthogonal, so $Q'\overline{Q'}D$ will project the observations D into the L-dimensional subspace within $\mathbb{C}^N$. If $L = \text{rank}\,(D)$, then the columns of D are already within the subspace so the projection does nothing. If, as in the example at the beginning of this chapter, we expect a signal of interest to lie in a low-dimensional subspace, and we further expect that component signals corresponding to our target will dominate, then this process can both estimate the subspace and reduce the noise.

While this example was motivated by thinking of D as a collection of observations, the result was a low-rank approximation of D, for which there isn't any special interpretation of the rows and columns. Rather than computing the eigendecomposition of the correlation matrix, we can use the singular value decomposition (SVD) of the data matrix D directly. The SVD gives $D = U\Sigma\overline{V}$, where the columns of U and V are orthonormal bases for the row and column spaces of D, respectively. Σ is a diagonal matrix containing the *singular values* of D, which are the square-roots of the eigenvalues of the correlation matrix (and are again assumed to be sorted in descending order). To perform a rank-L approximation in terms of the SVD, we define U' and V' as the matrices made from the first L columns of U and V, and Σ' as the $L \times L$ matrix with the top L singular values. The approximation is then $D' = U'\Sigma'\overline{V'}$.

Figure 34 shows the result of denoising a signal by subspace projection. Each observation of the signal is a sinusoid with random phase and amplitude, but the same (unknown) frequency. Because a sinusoid with arbitrary phase and amplitude can be expressed as a sum of sin and cosine terms, the observations lie in a rank-2 subspace. Because the frequency was not known *a priori*,

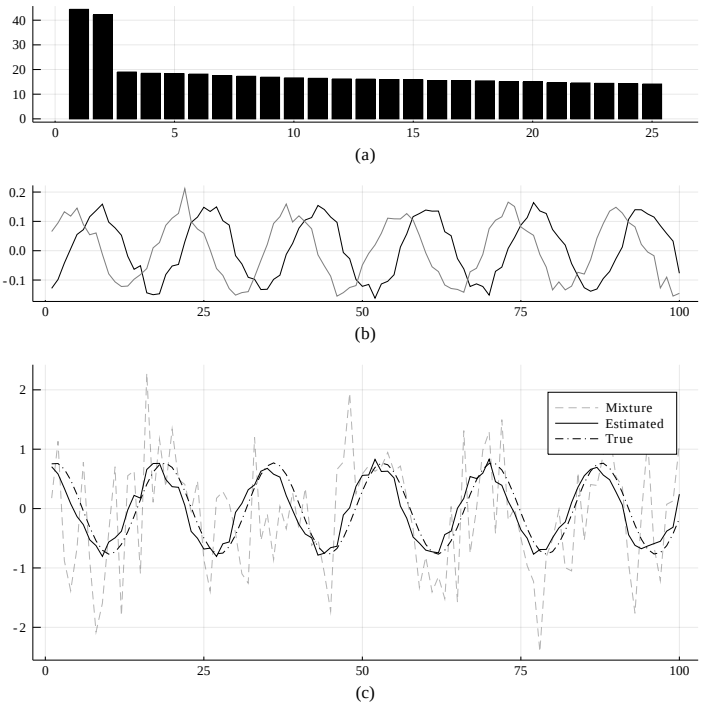

Figure 34: Subspace basis estimation on simulated data. The data consists of 200 observations of length-100 sinusoids, where each signal had a random amplitude and phase. The signals were mixed with white gaussian noise with average -6dB SNR. The subspace is estimated from the data using the singular value decomposition (chosen *a priori* to be rank-2). (a) shows the first 25 singular values, (b) shows the estimated subspace basis, and (c) shows an example before and after denoising.

the subspace was estimated from the data itself. Note that this projection will not completely remove the noise. One reason is that the subspace basis may not be estimated exactly. Additionally, some of the noise may be within the signal subspace by chance.

Applications in Multichannel Signal Processing

Recall the signal model that motivated the subspace approach:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,M} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N,1} & A_{N,2} & \cdots & A_{N,M} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \\ \vdots \\ n_N \end{bmatrix} \tag{23}$$

So far we have considered x, n, and the columns of A to be time-series, though there is nothing in the mathematics that has been specific to that interpretation. Instead we can consider a model widely-used in radio-frequency antenna array processing, where each vector is an instantaneous measurement of a complex-valued multichannel signal using an array with N sensors. Consider M uncorrelated narrowband (sinusoidal) sources in an anechoic environment. The spreading loss and time delay between each source and each microphone can then be encoded in the magnitude and

phase of a complex coefficient. This model is frequently used in farfield array processing, though in that context the magnitude differences between channels are often negligible. The ith column of A (denoted $A_{*,i}$) collects these coefficients together such that $w_i(t) = A_{*,i}(t)v_i(t)$ gives the contribution of v_i to x (also known as the *image* of v_i at time t).

This model gives rise to the classic beamforming approach to signal enhancement and localization - for a given source location or direction, source frequency, and array geometry, the corresponding column of A can be computed without reference to the data. We'll refer to this function as $a(\theta)$, where θ captures the relevant signal parameters. If the location of the source (represented by θ) is known, the array can be "focused" on it with $\widehat{v}(t) = \overline{a(\theta)}x(t)$. This can be understood as aligning all the channels of x so that they add coherently for a signal at the given location. For this reason the vectors $a(\theta)$ are often called *steering vectors*, because they aim the beam of the array. If the location is not known, beamforming can be used for localization by finding $\widehat{\theta}(t) = \text{argmax}_\theta \, |\overline{a(\theta)}x(t)|^2$, that is, finding the θ that maximizes the output power when it the array focused on that location. This is exactly the steered response power discussed in Part II.

Multiple Signal Classification (MUSIC)

Note that beamforming does not require (or take advantage of) statistical properties of the model. The Multiple Signal Classification (MUSIC) algorithm [53] starts by performing the eigendecomposition of the correlation matrix R_{XX}, as before. If the number of sources is not known it can often be estimated by inspecting the eigenvalues to find the dimensionality of the signal subspace. The intuition for this is that the steering vector for a given source is invariant, so the image of that source on the array is a set of scaled versions of the steering vector. Thus $x(t)$ is a linear combination of the steering vectors, with the source signals $v(t)$ providing the time-varying coefficients. From the eigendecomposition $Q\Lambda\overline{Q}$ and the rank estimate L we can define projections $Q_s\overline{Q_s}$ and $Q_n\overline{Q_n}$ Where $Q_s = Q_{*,1:L}$ and $Q_n = Q_{*,L+1:N}$, and their columns span the signal and noise subspaces, respectively. As mentioned earlier, the columns of Q_s are *not* in general the steering vectors corresponding to the sources - they are a unitary basis for the subspace that contains the steering vectors.

MUSIC takes advantage of the fact that the steering vectors $a(\theta)$ are constrained (by the array geometry) to a low-dimensional manifold within the larger $\mathbb{C}^N$ space. That is, often $\theta \in \mathbb{R}$, in the

[53] R. Schmidt (Mar. 1986). "Multiple Emitter Location and Signal Parameter Estimation"

case of 2D direction of arrival (DOA) estimation, or $\theta \in \mathbb{R}^2$ for 3D DOA (or 2D localization). Because the search space is small, $P(\theta)$ can often be searched exhaustively (or sampled on a fine grid).

MUSIC defines a metric $P(\theta) = \left|\overline{Q_n} a(\theta)\right|^{-2}$ which can be interpreted as the inverse energy of the projection of $a(\theta)$ into the noise subspace, or the inverse squared distance from $a(\theta)$ to the signal subspace. If the signal subspace was estimated accurately, there should be L points where the manifold $a(\theta)$ pierces through the subspace - at these points $P(\theta)$ goes to infinity.

So MUSIC takes advantage of the low-rank statistical structure of the linear model, and also the constraints that physics puts on the model. Of course this relies on having such a model that can be leveraged to constrain the signal subspace.

Nonnegative Matrix Factorization (NMF)

As seen in the previous sections, a linear mixing process can be represented as a matrix multiplication, and conversely recovering the sources and mixing coefficients can be framed as matrix factorization. Assuming the linear system is underconstrained, there are an infinite number of factorizations that are consistent with the observed signals, so the main research question then becomes which constraints lead to the most useful decomposition. Nonnegative Matrix Factorization decomposes a data matrix V into two matrices $V \approx WH$. Interpreting the columns of V as examples, the columns of W are basis vectors and H gives the mixture of basis vectors for each example (referred to as *encodings*). W and H are constrained to be nonnegative, so V is assumed to be nonnegative as well. NMF was proposed in the context of decomposing greyscale images of faces, where it was shown to successfully decompose the images into basis vectors representing variations on noses, eyes, etc. (the faces in the image dataset were white with dark features).

This approach has been widely-used in the audio source separation, and typically applied in the STFT domain. Audio data does not satisfy the nonnegativity assumption (the STFT is complex-valued), however it is common to model instead the total energy, under the assumption that for uncorrelated signals their energy adds linearly. Thus the mixture energy is modeled as a linear combination of the energy in each signal. Because NMF works entirely in the power domain and disregards phase, implementations generally re-use the mixture phase as the source phase, which is only accurate for time-frequency bins that are dominated almost entirely by one signal or the other. Signals for which this is

true are known as *w-disjoint orthogonal* [54], i.e. the supports of the signals' STFTs under window function w are disjoint sets. In fact the central problem of binary-mask-based source separation can be thought of as finding an invertible transform that takes the mixture signal into a space where the signals are disjoint orthogonal, masking each signal, then inverting the transform back into the time domain. The STFT is a popular choice because many signals (most notably speech) are approximately disjoint, however it is not the only choice.

[54] Alexander Jourjine, Scott Rickard, and Ozgur Yilmaz (2000). "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures"

Extensions to Convolutive Mixtures

NMF has also been applied in a multichannel convolutive context [55], though with stereo signals in a music source separation context. As in the approach proposed in the next chapter, they handle the convolutive mixture as a linear instantaneous mixture in each frequency band, with a complex-valued mixing matrix. They use NMF to further factorize the source matrix using NMF, which provides more consistency across frequency bands, and provides an extra constraint that solves the permutation problem inherent in other per-band source separation approaches such as frequency-domain independent component analysis (FD-ICA), where ICA is performed separately in each band. Using NMF as a source model also allows the source and mixing matrices to be identified individually.

[55] Alexey Ozerov and Cédric Févotte (2009). "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation"

Ozerov and others have introduced the concept of the *spatial covariance matrix*, where the image of each source in the multichannel mixture is considered a random vector, and the structure of its covariance matrix gives important insight into the mixing model [56]. They note that in the situation that the narrowband approximation holds, a convolutive mixture can be modeled with a rank-1 covariance matrix for each frequency band. When the narrowband approximation does not hold, reverberations of previous STFT frames are expressed as separate sources, increasing the rank of the covariance matrix.

[56] Ngoc QK Duong, Emmanuel Vincent, and Rémi Gribonval (2010). "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model"

Background: Signal Enhancement and Separation

Source separation and signal enhancement are closely-related and overlapping fields, but here we make a distinction that follows our definitions of interference and noise. Source Separation is primarily concerned with the problem of extracting multiple interfering signals. It is assumed that there are multiple signals of interest and that we can make strong assumptions about their structure (e.g. correlation in time, frequency, and/or space). In signal enhancement however, we are given a mixture of a target signal and noise - we can make only weak assumptions about the structure of the noise. Insofar as signal enhancement is in effect separating signal from noise, we generally focus on the properties of the signal to extract it, and the noise is the residual from the mixture after removing the target signal.

This work is primarily focused on signal enhancement - we seek to extract a target signal from the background noise. Obviously the ability to handle multiple simultaneous sources is desirable, so we point to extensions where the techniques explored here could be extended to a source separation context.

This chapter will cover several approaches, and provide examples from our ground truth dataset. They can be broadly categorized [57] as *spatially-oriented* and *source-oriented*. Spatially-oriented techniques focus on the inter-channel relationships, and generally make weak assumptions about the source features. Source-oriented techniques make use of stronger assumptions about the spectro-temporal structure of the sources, which can be determined *a priori* or learned from data. To take advantage of inter-channel spatial information, we will use low-rank filtering concepts introduced in the previous chapter.

[57] E. Cano et al. (Jan. 2019). "Musical Source Separation: An Introduction"

There has also been a variety of approaches under the guise of "speech enhancement", which can be thought of a form of separation, where the speech is separated from background noise or interfering signals.

One of the simplest is the classic Wiener Filter [58], which assumes a known signal and noise power spectrum and constructs a

[58] Philipos C. Loizou (2007). "Wiener Filtering". 1st

linear, time-invariant filter that admits each frequency to maximize how much of the signal is passed, filtering out the noise. Of course this leaves open the question of estimating the signal and noise spectrum. It also assumes a stationary signal. For nonstationary signals, the parameters of the Wiener filter can be updated across time, giving the adaptive Wiener Filter [59].

An alternative technique is Time-Frequency Masking [60], where the mixture is transformed into the time-frequency domain, multiplied by a (usually real-valued) mask, and transformed back into the time domain. The task then becomes estimating the best mask. Note that because multiplication in the frequency domain corresponds to convolution (filtering) in the time-domain, time-frequency masking is a form of adaptive filtering.

In recent work deep learning has become the dominant technique for source separation, whether estimating the time signal directly, estimating a time-frequency representation, or when using a mask. Ward et al. provide a good overview of the current state-of-the-art methods, including subjective comparisons with human listeners [61].

If a generative model is available for the target signal, one can also view source separation as a parameter estimation problem, where the goal is to use the mixture to estimate the parameters of the model, and resynthesize the target from scratch. This has the benefit that the "separation" is less susceptible to interference artifacts. This also provides a framework to further constrain the source estimates, as well as use the physics of the mixing process which are often well-understood. These can often be trained via expectation-maximization [62]. Some birds with simple calls might be well-modeled as amplitude- and frequency-modulated sinusoids, though other calls such as crows and geese have more complex harmonic structures that are less straightforward to estimate.

This work will focus primarily on the time-frequency domain, which provides a convenient framework for adaptive filtering, given some assumptions discussed in the following sections.

[59] Laurent Benaroya et al. (2003). "Non Negative Sparse Representation for Wiener Based Source Separation with a Single Sensor"

[60] Ozgur Yilmaz and Scott Rickard (2004). "Blind Separation of Speech Mixtures via Time-Frequency Masking"

[61] Dominic Ward et al. (2018). "SiSEC 2018: State of the Art in Musical Audio Source Separation - Subjective Selection of the Best Algorithm"

[62] Michael I. Mandel, Ron J. Weiss, and Daniel PW Ellis (2010). "Model-Based Expectation-Maximization Source Separation and Localization"

Spectral Subtraction

Spectral subtraction is a simple technique for enhancing a target signal in a noisy mixture in the time-frequency domain. It relies on the observation that when the noise is uncorrelated with the target, the energy of the mixture is the sum of the energies of the noise and target. If a good estimate of the noise energy is available, then the signal energy can be estimated by simple subtraction. When

first proposed, this was implemented by computing the Fourier transform of each frame, subtracting the noise power from the mixture power, then converting back into the time domain via the inverse Fourier transform, using the phase from the original mixture [63]. Frequency bins that would have negative power were set to zero. Shortly thereafter Berouti et al. [64] extended the technique by overestimating the noise power by a tunable factor α, and setting the floor based on the noise estimate for each frequency, rather than letting the power go to zero. These additions were primarily intended to address the issue of *musical noise*, which a common issue in spectral-subtraction where subtracting the average power leaves behind "islands" in the time-frequency representation that are audible as warbling tonal artifacts. Both these implementations relied on a speech detector to estimate periods of "silence", which were used to estimate the noise power spectrum.

[63] Steven Boll (1979). "Suppression of Acoustic Noise in Speech Using Spectral Subtraction"

[64] Michael Berouti, Richard Schwartz, and John Makhoul (1979). "Enhancement of Speech Corrupted by Acoustic Noise"

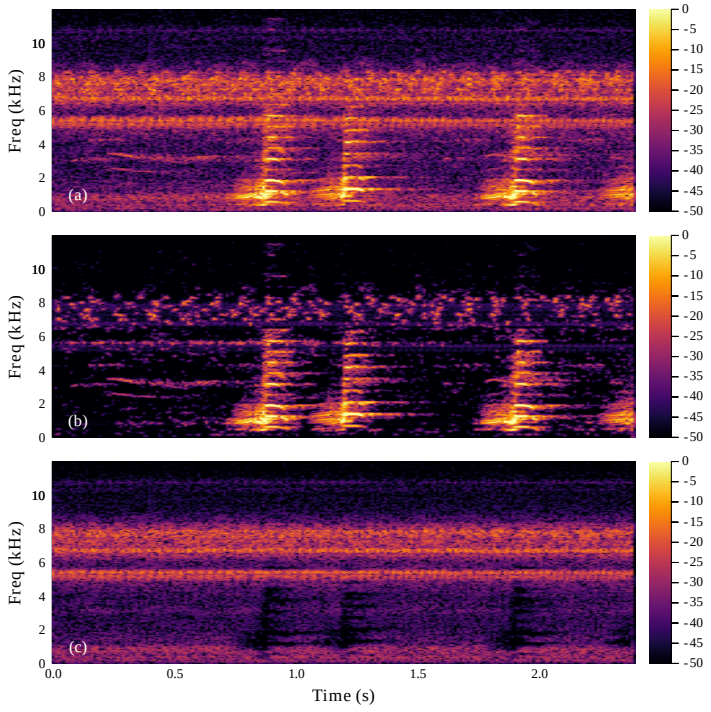

Figure 35: Spectrogram plots showing the spectral subtraction method. The three prominent harmonic bursts are the target sound (a goose call). Insect noise dominates the 6-8kHz band, and the narrowband signals in the first 0.75s are an interfering bird call. (a) is the original audio, (b) is the result of spectral subtraction, and (c) is the residual noise. This example was generated with $\alpha = 6, \beta = 0.01$. Noise PSD was estimated for each band as the minimum power after smoothing with a 250ms gaussian window.

Later Martin investigated an alternate approach to noise PSD estimation by observing that within each frequency band, the signal frequently reverts to the noise floor, so tracking the minimum energy over a window gives an estimate for the noise floor over that window [65]. In practice it is necessary to smooth the energy over time to reduce the variance of the estimate. Note that by definition this will be an underestimate of the noise floor (the minimum of a set of samples will be below the mean). Martin proposed es-

[65] Rainer Martin (1994). "Spectral Subtraction Based on Minimum Statistics"

timators for the degree of underestimating, but here we simply compensate by hand-tuning α, the oversubtraction parameter.

We implemented a similar scheme, assuming the noise power spectrum was constant over a given sample of length 2.5s. The spectrograms in 35 show the results. These were generated with an aggressive α of 10, which helps substantially with musical noise, but would likely cause audible artifacts in lower SNR mixtures. Notice that most of the broadband noise has been effectively removed, but the insect noise in the 6-8kHz band has left substantial musical noise (visible as isolated dots on the spectrogram), because the energy is highly variable in those bands. Also notice the shadows left by target sound in the residual - they are audible as brief dips in the overall energy. Also note that because spectral subtraction can only remove stationary noise, it doesn't differentiate between target and interfering sounds, as shown by the interfering bird call in the first 0.75s. The traces in the first 0.75s are not part of the target.

Wiener Filtering

Given a known *random process* x and a target signal y, both of which are zero-mean and jointly *wide-sense stationary* (WSS) (their mean, autocorrelation, and covariance are not a function of time), the Wiener filter h is the filter that minimizes $\mathbb{E}\left[|h * x - y|^2\right]$, and is given by $\widetilde{h}(\omega) = \frac{S_{xy}(\omega)}{S_{xx}(\omega)}$, where S_{xy} is the *cross-spectral density* (CSD) of y and x, defined as $S_{xy}(\omega) \triangleq \mathbb{E}\left[\overline{\widetilde{x}(\omega)}\widetilde{y}(\omega)\right]$, and S_{xx} is known as the *power spectral density* PSD) of x[66]. Note that this definition also applies to deterministic signals, where the expectation just provides the signal itself.

[66] Alan V. Oppenheim and George C. Verghese (2015). "Wiener Filtering"

One application of the Wiener Filter is known as Wiener Deconvolution, where an observed noisy signal $x = b * s + n$ is assumed to be generated from a target signal s convolved with a known impulse response b, with additive noise n. The Wiener filter h minimizes $\epsilon = |\hat{s} - s|^2 = |h * x - s|^2$. This may at first not seem very useful because the Wiener framework assumes we know the target signal. However, under the assumption that s and n are uncorrelated, the filter is computed as:

$$\widetilde{h}(\omega) = \frac{\overline{\widetilde{b}(\omega)} S_{ss}(\omega)}{S_{bb}(\omega) S_{ss}(\omega) + S_{nn}(\omega)} \tag{24}$$

Notice that this does not require the Fourier transform of s and n, only their power spectral densities, which can often be estimated. Intuitively this filter corresponds to frequency-

domain division (decorrelation) for frequencies with high SNR ($S_{ss}(\omega) \gg S_{nn}(\omega)$), and goes to zero for frequencies with low SNR ($S_{ss}(\omega) \ll S_{nn}(\omega)$), which avoids the instability introduced by naive frequency-domain division in bins where the noise is large but the transfer function is small. Wiener Deconvolution is also used for transfer function identification (or impulse response estimation) in situations where the source signal is observable but not controllable, e.g. in a teleconferencing application.

In the source separation context the same framework is also often used, including the assumption that the signal and noise are uncorrelated. The signal model generally does not include a transfer function to be deconvolved however, which simplifies equation 24 to

$$\widetilde{h}(\omega) = \frac{S_{ss}(\omega)}{S_{ss}(\omega) + S_{nn}(\omega)} = \frac{S_{ss}(\omega)}{S_{xx}(\omega)} \tag{25}$$

Because the power spectral density is real-valued, h is a zero-phase filter. Intuitively it can be interpreted as a weighting factor or soft-mask applied to each frequency band, where high-SNR bands are allowed through, and low-SNR bands are attenuated.

Recall that this analysis assumed that s and n were WSS, an assumption that is not applicable for source separation applications. In this context it is common to apply the Wiener filter adaptively to a succession of time windows in the STFT domain. In our case we have assumed the noise PSD to be stationary, so the filter becomes

$$\widetilde{h}(\omega, t) = \frac{S_{ss}(\omega, t)}{S_{ss}(\omega, t) + S_{nn}(\omega)} = \frac{S_{ss}(\omega, t)}{S_{xx}(\omega, t)} \tag{26}$$

While the Wiener Filter provides the optimal linear filter for extracting one signal from another (in the MSE sense), it relies on the signal and noise PSD, which need to be estimated. The minimum-smoothed-power estimator described in the previous section provides a noise PSD estimate $\hat{S_{nn}}(\omega)$ and spectral subtraction provides a signal PSD estimate $\hat{S_{ss}}(\omega, t) = S_{xx}(\omega, t) - \hat{S_{nn}}(\omega)$. To simplify notation we ignore the case where $S_{xx}(\omega, t) < \hat{S_{nn}}(\omega)$, in which case the estimated signal and the filter coefficient are both zero. This generates the Wiener filter described by Equation 27.

$$\begin{aligned}\widetilde{h}_{wiener}(\omega,t) &= \frac{\hat{S_{ss}}(\omega,t)}{S_{xx}(\omega,t)} \\ &= \frac{S_{xx}(\omega,t) - \hat{S_{nn}}(\omega)}{S_{xx}(\omega,t)} \\ &= 1 - \frac{\hat{S_{nn}}(\omega)}{S_{xx}(\omega,t)} \end{aligned} \tag{27}$$

It's informative to compare the results of applying spectral subtraction directly (as in the previous section) vs. using the Wiener framework.

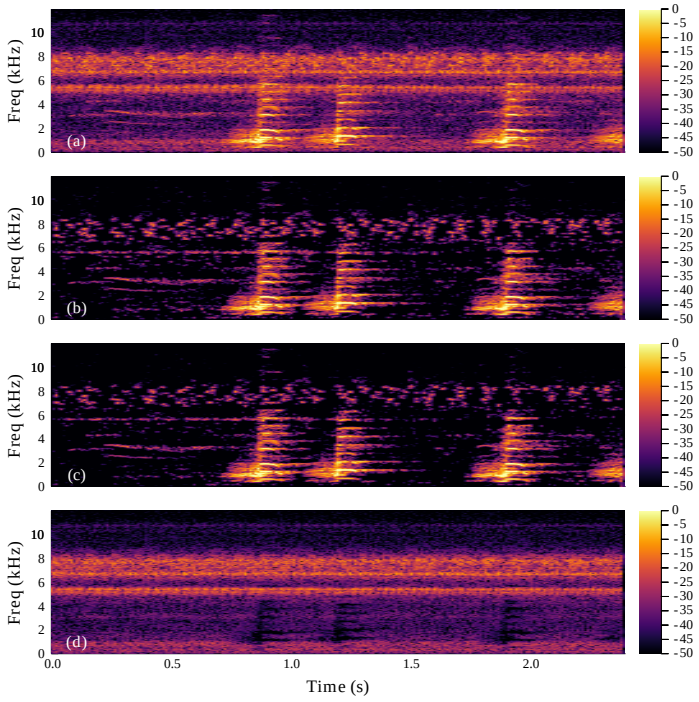

Figure 36: Spectrogram plots showing the Wiener Filter method compared with spectral subtraction. (a) is the original audio, (b) is the result of spectral subtraction, (c) is the result of Wiener filtering and (d) is the residual noise. This example was generated with $\alpha = 6, \beta = 0.0$. Noise PSD was estimated for each band as the minimum power after smoothing with a 250ms gaussian window.

Both methods essentially modify the STFT magnitude on a bin-by-bin basis, and can be thought of as zero-phase time-varying filters. Spectral subtraction isn't typically described as a filter, but we can extract the equivalent filter by dividing the estimated signal by the observed input.

$$\begin{aligned}\widetilde{h}_{specsub}(\omega,t) &= \frac{\hat{\widetilde{s}}(\omega,t)}{\widetilde{x}(\omega,t)} = \frac{\left|\hat{\widetilde{s}}(\omega,t)\right|}{\left|\widetilde{x}(\omega,t)\right|} \\ &= \frac{\sqrt{\hat{S_{ss}}(\omega,t)}}{\sqrt{S_{xx}(\omega,t)}} \\ &= \frac{\sqrt{S_{xx}(\omega,t) - \hat{S_{nn}}(\omega)}}{\sqrt{S_{xx}(\omega,t)}} \\ &= \sqrt{1 - \frac{\hat{S_{nn}}(\omega)}{S_{xx}(\omega,t)}} \end{aligned} \tag{28}$$

This demonstrates that the filter implicit in the spectral subtraction process, as well as the Wiener filter using the same power estimates, can both be expressed in terms of just the observed signal and the noise PSD estimate. Additionally, we see that the spectral subtraction filter is the square root of the Wiener filter. Recent work [67] has investigated this relationship, as well as the more general *parameterized Wiener filter* where the power of the filter is included as a continuous parameter.

[67] Mathieu Fontaine et al. (2017). "Explaining the Parameterized Wiener Filter with Alpha-Stable Processes"

Figure 36 shows the result of the Wiener filter on the same audio example, and offers a comparison with direct spectral subtraction, using the same parameters. Due to the squaring of the filter coefficients, the Wiener filter shows slightly more noise suppression, but perceptually the examples are difficult to distinguish.

An Aside on STFT Consistency

It is common in STFT processing to overlap the adjacent windows, to maintain information that could be lost due to windowing, and also so that adjacent frames are cross-faded on resynthesis, reducing windowing artifacts. Because the STFT is overcomplete, there are STFT signals that don't correspond to any time-domain signals. These STFTs are known as *inconsistent* [68].

[68] Jonathan Le Roux and Emmanuel Vincent (Mar. 2013). "Consistent Wiener Filtering for Audio Source Separation"

Because the STFT is a linear transform, the set of *consistent* STFTs lies on an N-dimensional subspace (where N is the length of the signal). When a signal is modified in the STFT domain, it is likely that the modified signal is no longer consistent. Performing the ISTFT via overlap-add generates a time-domain signal, but bringing the signal back into the STFT domain necessarily will not give the same result as the modified STFT-domain signal.

As a concrete example, consider a synthetic STFT that is all zeros with a single bin equal to one. Performing the ISTFT will

generate a single ISTFT basis function (a modulated window). Performing the STFT on that modulated window will have multiple non-negative bins in the neighborhood of the original impulse, due to the overlap of the windows and spectral leakage between adjacent frequency bins.

Spectral subtraction provides an additional example, as seen in Figure 37. The STFT produced by the spectral subtraction process is not constrained to be consistent, so after a round-trip through the time domain the signal is not maintained.

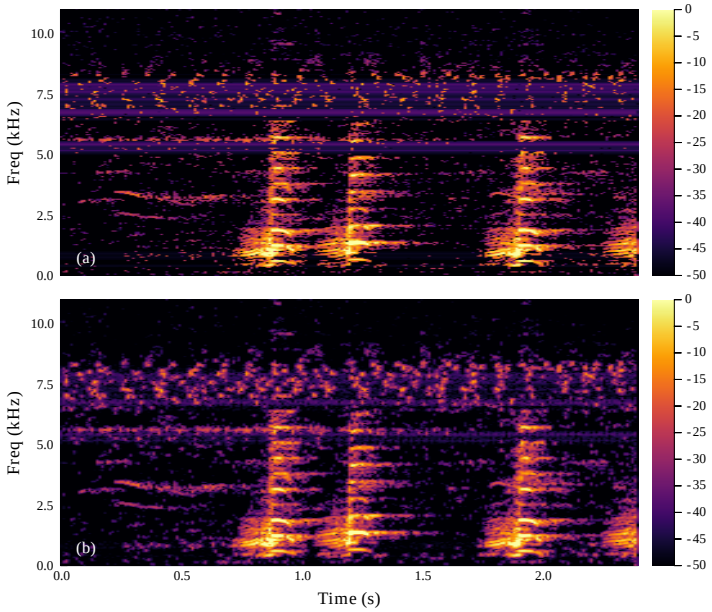

Figure 37: Spectrogram of spectral subtraction before and after the inverse STFT. (a) Shows the STFT-domain signal after spectral subtraction. (b) shows the same signal after the ISTFT is performed (and another STFT performed for display). Parameters are the same as 35.

A Low-Rank Filter for Foreground Separation

The methods described in the previous chapters present various approaches to this problem in a multichannel context, but there are assumptions made that do not hold true for our problem, and require modifications. In this chapter we define the assumed signal model, and propose a method for separating spatially-compact foreground sounds from spatially-diffuse background sounds.

Signal Model

Let $\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}$ be observed microphone signals, also called the *mixtures*, and $\mathbf{s_1}, \mathbf{s_2}, ..., \mathbf{s_M}$ be sources, with locations given by $p_1, p_2, ..., p_M$. Boldface variables are time-domain signals that are considered to be monophonic and omnidirectional. The signal model is defined as:

$$\mathbf{x_i} = \sum_{j=1}^{M} \mathbf{h_{ji}} * \mathbf{s_j} + \mathbf{b_i} \tag{29}$$

Where $\mathbf{h_{ji}}$ is the transfer function of source j to microphone i, and $\mathbf{b_i}$ is background noise which we assume to be independent at each microphone.

Within the scope of this work we take the first source $\mathbf{s_1}$ as the *target*, with any other sources considered *interferers*. Further, we do not model the interferers separately and instead include them in the background noise. The goal of the system is to estimate the target signal, its position, and a background signal $\mathbf{b_i}$ at each microphone position, given the observed mixtures.

The *image* of a source is the signal at the microphone due to that source, i.e. $\mathbf{h_{ji}} * \mathbf{s_j}$. The background at each microphone is in effect the residual that is left once the image is removed. Once these parameters have been estimated, they can be used as inputs to the spatial audio renderer. The target audio becomes a point source that can be placed at its true location, and the background sounds can be placed as diffuse sources at each microphone location.

Handling Large Inter-Microphone Distances

In many sensor array applications, the propagation delay being estimated is small (relative to any periodicity in the signal) because the sensors are close together. In this case the delay can be measured independently in each frequency component and in narrowband applications often only one frequency is considered. The delay is then given by $\frac{\Delta\phi}{\omega}$, where $\Delta\phi$ is the phase difference in radians, and ω is the frequency in radians/s. If the delays are longer than the period in question, the phase wraps around and becomes ambiguous. This case is often called *spatial aliasing*. In our case the delays are orders of magnitude longer than the periods of the signals of interest, so narrowband delay estimation is impossible.

In the presence of spatial aliasing, estimating a unique delay requires wideband signals, where delay in the time domain corresponds to a linear phase shift in the Fourier domain. Fortunately the signals of interest in this work are wideband (though they often have strong periodic components), and thus the delay estimation can be performed by cross-correlation. In cross-correlation-based time delay estimation, strong periodic components in the signals cause large oscillations in the cross-correlation function. These oscillations make the task of identifying the cross-correlation peak more difficult, but can be largely addressed by employing GCC-PHAT, as discussed in Part II.

The other challenge presented by long inter-microphone delays lies in the comparisons we would like to make between channels to enable separation. Recall that subspace techniques like MUSIC require the energy to be correlated between the channels, so the window under analysis needs to allow for any delay. Longer windows require more computation to process, and if the window is long relative to the duration of the signal, much of the window is noise. Additionally, for statistical estimations we assume the properties to be estimated are stationary over the course of the measurement. If they are not stationary, they can often be approximated as such, but longer windows degrade the approximation.

To account for this we assume the source-to-microphone transfer function can be decomposed into a bulk delay and a zero-phase filter. We use the localization estimate described in Part II to align the signals, so the analysis window size can be chosen independently of the delay times.

Errors in the estimated source location (generally less than 4m), measured microphone locations (generally less than 1m), or the speed of sound (which is somewhat affected by temperature

and wind) can cause the location-based alignment to be sub-optimal. The alignment is improved by following the location-based alignment with a fine-tuning that maximizes the cross-correlation within a $\pm$ 6ms lag window.

Full Per-Channel Transfer Functions

The work reviewed so far considered a mixing model where the contribution of each source to each observed signal could be represented with a complex multiplication. In this work, the signals of interest are wideband and the source is assumed to have a different transfer function to each microphone. To this end, we use the Short-Time Fourier Transform (STFT) to implement a filterbank, and performed the subspace de-noising on each band individually. The STFT was performed using a 1024-sample FFT and 512-sample hop size. The audio was sampled at 48kHz, so this corresponds to a 21ms window and 11ms hop. Each band was demodulated to center it at 0Hz. The analysis was performed on 3-second samples, matching the length used for the localization experiments.

If we consider X_k to be the multichannel output of the filterbank at the kth band, with a single source the linear system becomes:

$$\begin{bmatrix} | \\ X_k(t) \\ | \end{bmatrix} = \begin{bmatrix} | \\ A_k \\ | \end{bmatrix} v_k(t) + \begin{bmatrix} | \\ n_k(t) \\ | \end{bmatrix} \tag{30}$$

As in the previous work, A_k is a complex-valued vector that contains the mixing coefficients for the source into each microphone. Assuming the source is stationary and the transfer function is otherwise time-invariant, A_k should be constant. v_k is the kth band of the source.

Note that here we are making the frequently-used *narrowband approximation* within each frequency band, which is the assumption that the effect of an LTI system with transfer function $\widetilde{h}$ on a signal with Fourier transform $\widetilde{x}$ can be captured by a single complex coefficient $\widetilde{h}(\omega_0)\widetilde{x}$. That is, $\widetilde{h}(\omega)\widetilde{x}(\omega) \approx \widetilde{h}(\omega_0)\widetilde{x}(\omega)$, which is true for a sinusoid with frequency ω_0, and approximately true for narrowband signals with energy concentrated near ω_0. The extent to which the narrowband approximation is appropriate depends on the bandwidth of the signal relative to how quickly the transfer function $\widetilde{h}(\omega_0)$ is varying near ω_0. That is, the approximation assumes a narrowband signal and "smooth" transfer function.

Note that this definition has a dual in the time domain, given by taking the inverse Fourier transform of both sides: $(h * x)(t) \approx$

$\widetilde{h}(\omega_0)x(t)$. For a stationary (complex) sinusoid x this is satisfied by any h because any linear combination gives another sinusoid with an amplitude and phase shift. For an approximately narrowband signal this tells us that the length of the impulse response must be short relative to the time scale of the modulations.

Notice that in both the time and frequency domain the approximation gives a multiplication by $\widetilde{h}(\omega_0)$. This duality is important in the context of the STFT because each point in the time-frequency plane can be thought of as part of the spectrum of a windowed signal (frequency domain), and also as the output of a filter in a filterbank (time-domain). The duality of the narrowband approximation tells us that both perspectives are equivalent in the signal's response to a linear system.

Because we are assuming a single source, the first eigenvector of the correlation matrix of X_k gives A_k. Because we are estimating the covariance from the observed data, it is helpful to think in terms of the SVD. Consider a $N \times T$ data matrix D_k, where N is the number of microphone channels (12 or 13 for our experiments) and T is the number of samples from the filterbank (equivalently the number of frames in the STFT, on the order of 100-150 for our experiments).

Given the observations of X_k, we can only estimate A_k and v_k up to a complex scalar factor - that is, their amplitude and phase are free to vary inversely while still being consistent with the model. This means that while the model is useful for removing noise via subspace projection, it does not estimate the source and transfer function individually, and the output of the system is a multichannel signal, not a source estimate. Thus, the question remains of how to choose which output channel to use as the source for resynthesis. In this work we simply use the channel from the microphone closest to the source.

Other methods like frequency-domain ICA do attempt to extract the source, but still work with each frequency individually. In that case there remains a permutation problem, where the sources extracted from each frequency band need to be grouped together.

Additionally, notice that the output at time t only depends on the input at time t. While this model can capture short-duration transfer functions such as the filtering caused by air absorption, it does not model longer-duration effects such as reflections. This means that for the output $X_k(t)$, contributions due to earlier values of v_k are treated as noise.

In testing, we noticed that the target signal often appeared strongly in the second eigenvector as well, likely due to strong reflected components. We considered using a rank-2 estimate

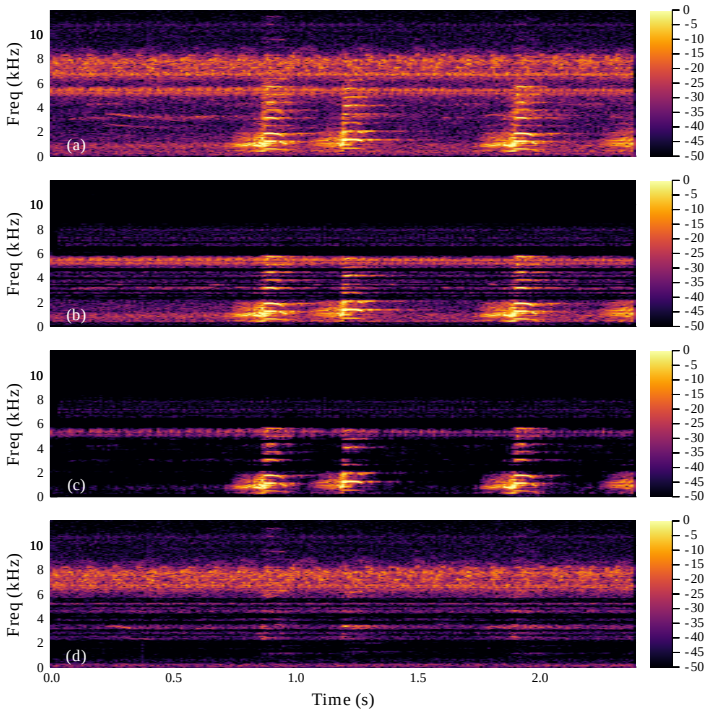

Figure 38: Spectrogram plots showing the rank-1 filtering method. (a) is the original audio, (b) is the result of the rank-1 filtering with unweighted SVD, (c) is with the iterative weighting scheme and (d) is the residual noise from the weighted version.

instead, however the additional signal energy was offset by additional noise as well. For generating the background residual, we thus skipped the second eigenvector when defining the noise subspace, and instead used the third through the last. This helped reduce the amount of target in the background substantially.

It's important to note that if our signal subspace were higher dimensional (i.e. it contains linear combinations of multiple sources) then this method can identify the subspace, but not recover the original signals. The basis vectors given by PCA will be orthogonal to each other (the eigenvectors give a unitary basis), which is of course not in general not true about the original basis. To recover the original basis signals we need to be able to apply some other constraints based on *a priori* knowledge of the signals. As mentioned before, the MUSIC algorithm is one widely-used method that can be applied when the basis functions are known to lie on some lower-dimensional manifold. To apply this technique in our context would require much more precise knowledge of the microphone positions (which could be achieved through calibration), though the results may be corrupted by time-varying properties like wind. To characterize the low-dimensional manifold we'd also need an accurate model of the transfer functions from each source location to each microphone location.

Weighting

If there are non-gaussian interfering signals present in the observed data, the rank-1 approximation will sometimes generate a result that splits the difference between the target and the interference. To reduce this effect, an iterative weighting scheme performs the rank-1 filtering on all the bands, and then assigns a weight for each time-frequency bin, based on the fraction of the total energy for that bin that is within the signal subspace. The weights are smoothed across frequency with a 31-point gaussian window, then used for another rank-1 filtering process. This is iterated until the weight matrix converges within a tolerance (typically 2-3 iterations). The frequency-smoothing is based on the observations that energy in adjacent bins is often correlated, so high-energy bins that were correctly identified by the rank-1 filtering can add extra weight to lower-energy bins. Additionally this helps counteract some of the discontinuities across frequency that occur because each band is processed independently.

Figure 38 shows the result of rank-1 filtering. With no weighting we see very little noise reduction in bands with significant target energy, creating horizontal bands. This is likely because the signal subspace is not being estimated accurately. The iterative weighting scheme improves inter-peak noise reduction substantially. Relative to the spectral subtraction and Wiener filtering versions we can see less reverberation and also very little musical noise. The residual has lost a lot of energy in the bands dominated by the target, even where the de-noised target doesn't seem to be present. This is likely due to energy within the second principle component, which includes both target and background energy, and is not included in either the signal or noise subspaces.

Results

I evaluated the proposed Rank-1 filter on our dataset of naturalistic recordings, along with two baseline methods. The methods we compared were spectral subtraction, Wiener filtering with smoothing, rank-1 filtering, and rank-1 filtering with the iterative re-weighting scheme. The results are shown in Figure 39. Each datapoint included in the violin plot is one channel of one recording. There are 201 multichannel recordings, split into 787 3s segments with a 1.5s hop size, giving 787 segments. Each segment has 12 or 13 channels. Processing the combined localization and enhancement took several hours on a dual-core laptop with 2.7GHz Intel Core i7 processor. To avoid confounding the results with errors due to localization, the ground-truth location was used for the Rank-1 methods, which require the signals to be pre-aligned.

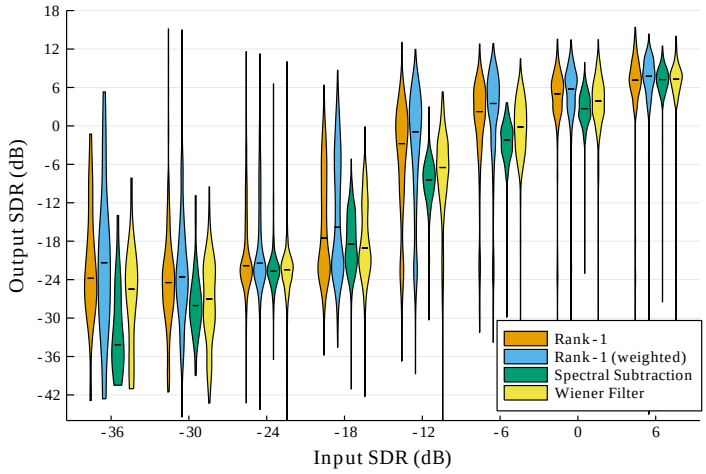

Figure 39: SDR Improvement by performing signal enhancement on our dataset of naturalistic recordings, using four techniques: spectral subtraction, Wiener filtering (using spectral subtraction for the signal and noise estimates), rank-1 filtering, and rank-1 filtering with weighting. All methods used a 1024-point STFT with 512-point hop size, and cosine windows for both analysis and synthesis. Noise PSD estimation for the spectral subtraction and Wiener filtering both used the minimum power in each band after smoothing with a 0.25s gaussian window. Both also used an oversubtraction factor of four, with no minimum noise floor. The Wiener filter was smoothed in time and frequency with an 11-point window. Rank-1 filtering was performed separately on each band. The input SDR is quantized in 6dB increments for display.

We see that the weighted rank-1 filtering performed best across all input samples. The dramatic improvements at very low input SDR are somewhat surprising. One consideration is that with rank-1 filtering all channels in the recording are denoised jointly, so the higher-SDR channels can help improve the lower-SDR ones. However, this does not explain the performance of the Wiener

filter, in which each channel is processed separately. It is possible that this is due to an artifact of the SDR metric - because it admits the target with any 512-point FIR filter applied, it is possible that these methods are somehow creating an easier signal for the SDR estimator to find [69]. Another observation is that all the methods seem to approach odB improvement when the input SDR approaches 6dB, implying that there is perhaps little improvement to be had because the signal is already relatively clean.

[69] Jonathan Le Roux et al. (Nov. 2018). "SDR - Half-Baked or Well Done?" arXiv: 1811.02508 [cs, eess]

Limitations and Future Work

Multiple Sources We currently just look for the highest-energy source within each band and attempt to enhance that signal. This has a several issues:

- If there are multiple strong sources within a given frequency band the SVD will try to split the difference. To resolve this we'd need to include a partitioning step that assigns each bins to its respective source.
- If sources are in disjoint sets of bands, the algorithm will enhance one source in some bands and a different source in other bands. Resolving this requires doing more analysis across bands, with better source modeling and/or with constraints on the transfer functions implied by the per-frequency spatial subspace basis.

Moving Sources As with the proposed localization method, this approach to separation assumes the target source is stationary. The spatial covariance matrix is estimated assuming it is time-invariant within the analysis window, so moving sources will degrade separation performance. This assumption could be relaxed to allow the covariance matrix to change over time, though tracking these changes would require a more sophisticated estimation technique.

Correlations over longer time windows One major area where our model does not fit reality is that we treat each STFT frame as independent. Given the decay times observed in the impulse response survey that is clearly not the case. One simple adaptation would be to experiment with longer STFT windows. Another approach would be to try to estimate the actual impulse response from the data and account for the energy in previous frames explicitly. Transfer function estimation is a well-studied problem for acoustic echo cancellation for videoconferencing. In that context, you have access to both the input and the convolved output, but similar techniques should be applicable in our context where we have multiple samples of the output convolved with different

transfer functions.

Source Signal Estimation Because of the nature of the low-rank factorization, the magnitude of the phase of the source and transfer function are not estimated independently, so they can vary inversely and still be consistent with the observed signal. Prior work[70] has demonstrated the value of combining a more opinionated source model with the physically-motivated mixture model. Machine learning could be applied to learn constraints on one or the other (or both) from the data, which would allow the actual source signal to be estimated, rather than the image of the source on each microphone.

[70] Alexey Ozerov and Cédric Févotte (2009). "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation"

Event Detection Currently the system does not have an explicit event detection step, so some of the low-SDR results could be due to cases where there is a gap in the sample that is longer than the analysis frame. Results could be improved by handling this explicitly. Within a practical soundscape resynthesis application, event detection would be useful to decide when a point source should be created by the rendering system. Recent work[71] has shown successes in using a multichannel coherence metric in a wildlife localization context.

[71] Matthew Wijers et al. (Nov. 2019). "CARACAL: A Versatile Passive Acoustic Monitoring Tool for Wildlife Research and Conservation"

Test On Ogg-Encoded Data Currently the system has been tested with raw PCM audio recorded with known source signals. We have a large dataset of audio encoded in ogg-opus format, which the system should be validated on as well. Lossy encoding may affect the phase at higher frequencies, where human perception is not very sensitive to it. That gives limited bandwidth which would affect localization accuracy. It also might affect the signal enhancement, though only the higher bands should be impacted.

Test on Degraded Array More research is needed to characterize the effect of microphone failures leading to a sparser array. One way it might affect results would be less accurate target subspace estimation, because there would be fewer example of the target present. Because the noise suppression is based on subspace projection, we would also expect to see a reduced ability to remove noise, as the target subspace would be more aligned to the observed channel vectors, so less noise would be projected out.

Part IV

Rendering the Soundscape

Demo Application

To test the proposed resynthesis approach and demonstrate the end-to-end functionality of the system, I developed a 3D application that can be run on a standard personal computer. The user is first given some context and introduced to the proposed system. They are then free to explore a 3D representation of the monitored area of Tidmarsh, from a first-person perspective. As the user explores, a soundscape is rendered to headphones, reflecting their location and orientation in the virtual world.

The application was built within the Unity engine, a development environment which is frequently used for creating video games. The engine provides functionality for creating a virtual world and placing objects within it, as well as defining behavior for those objects in code. In the context of virtual reality (VR) and augmented reality (AR) applications, the auditory scene is made up of auditory objects, or sources - objects in the environment that produce sound (and do not necessarily have a visual manifestation in the virtual world.

The audio is spatialized using the Resonance Audio SDK[72]. In the context of virtual reality (VR) and augmented reality (AR) applications, the function of the spatializer is to process a stream of audio from each source in the scene, along with parameters for each source, and render a playback stream for the listener. Different spatializers support a variety of parameters for each source, and often provide different source types. The most basic is an omnidirectional point source, which has a location in virtual space, and emits sound equally in all directions. Depending on the spatializer, the sound designer may also be to specify a source's "size". When the user is far from the source relative to its size, the listener hears the source from a well-defined direction. When the user is near the source it becomes less strongly directional. For example, a swarm of insects may be audible from a particular direction from afar, but when one is within it, it is heard from all directions.

The spatializer renders the auditory scene based on the sources'

[72] Marcin Gorzel et al. (Mar. 2019). "Efficient Encoding and Decoding of Binaural Sound with Resonance Audio"

relative positions to the listener, taking into account both orientation (a source to the listener's right should be perceived as such) and distance (near sounds are louder than far ones because of spreading losses). Some spatializers (including Resonance) can also incorporate the geometry of the environment, including reflections and reverberation.

In both the baseline and separated conditions a red sphere is created at the target location as a visual indicator, as seen in Figure 40. This allows users to evaluate whether the perceived location matched the actual location.

The output from the spatializer typically has as many channels as the user's hardware, i.e. a 5.1 "surround sound" system would require six channels, and a listener wearing headphones would require two.

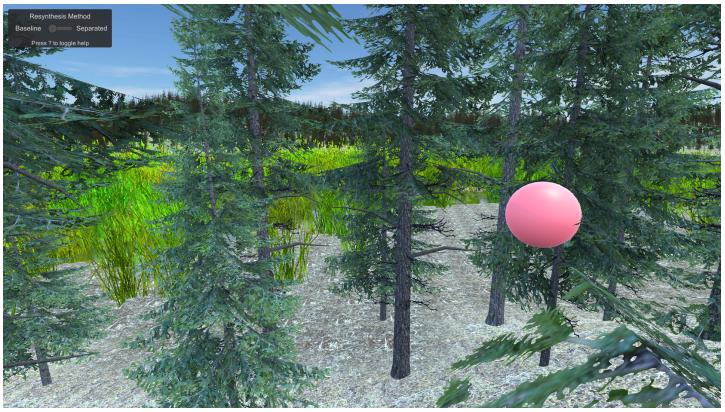

Figure 40: Rendering of the target location in the demo application.

Resynthesis Approach

The source material comes from our naturalistic sample recordings, where we played 12 different audio clips from several locations on site, using a portable speaker as described in Part I. The audio is stored on disk as Ogg Vorbis files, and distributed with the application.

The listener can alternate between two different approaches to rendering the soundscape, which are diagrammed in Figure 41. The first is a baseline, where the signal from each microphone is placed as a sound source at the microphone's location, as well as a model of the deployed microphones, as seen in Figure 42. In this configuration the microphone becomes a sort of virtual speaker. As mentioned in the motivation for this work, this limits the ability to accurately perceive the source's location, and creates undesirable echoes. In this implementation the size of each source is set such that there is generally one source that the user is "within" and the rest are heard directionally. This provides a continuous soundfield that varies with location, but doesn't give the perception that diffuse noises (such as wind or insects) are coming from a particular point in space.

The proposed approach is implemented using audio rendered off-line through the rank-1 filtering approach, with weighting. This outputs two channels per microphone - one with the target signal and one with the residual. Similar to the baseline, we place a source at each microphone location, though in the proposed approach we are using the residual, so very little energy from the target signal should be present. These residual sources are sized as in the baseline, creating a diffuse background soundscape. The system renders the target as a point source at the true target location. As a proxy for the source signal we use the de-noised target signal from the closest microphone, which also reduces the reverberation. The target location comes from the ground-truth data, so we are testing the resynthesis independently of the localization accuracy.

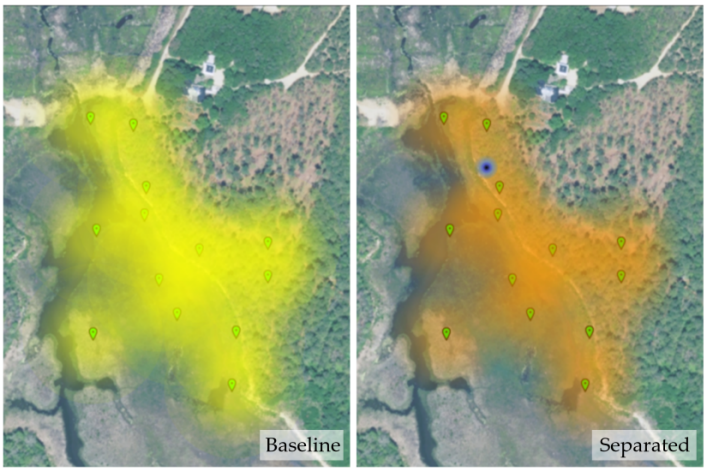

Figure 41: Diagram of the resynthesis approaches compared in the demo application. On the left is the baseline condition, where each microphone is represented as a diffuse source. When the user is far from these sources they are each perceived directionally, but when nearby they are rendered monophonically. On the right is the separated condition. The background and foreground are first separated with the approach descriped in Part III. The background signals for each microphone are rendered diffusely as in the baseline, but the foreground source is rendered as a point source at the actual location. Because we do not have access to the original source signal, the denoised signal from the closest microphone to the source is used as a proxy.

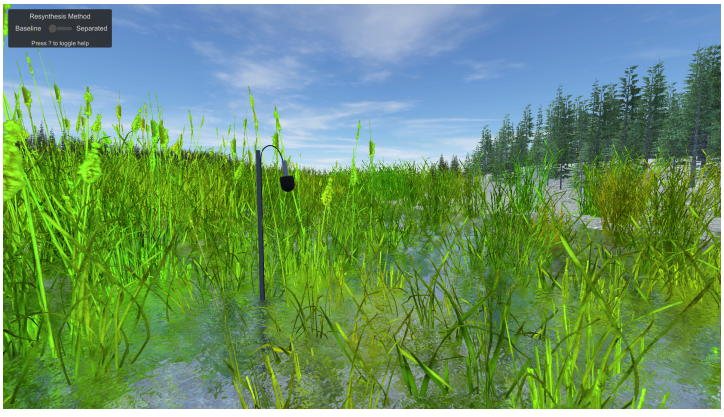

Figure 42: Rendering of a microphone in the demo application.

Results

To evaluate the demo application, download links were posted to a variety of mailing lists, primarily focused on audio technology, auditory perception, and bioacoustics, as well as a list of MIT Media Lab students, faculty, and alumni.

Users were asked to fill out a survey after the experience, which is included in *AppendixA*. They were instructed to use headphones, and asked to only fill out the survey if they had been able to do so. 47 people tried the application and responded to the survey, and self-described their experience working with sound as "none" (17, 36.2%), "audio hobbyist" (12, 25.5%), or "audio professional" (18, 38.3%). All but one participant reported being able to differentiate between the two methods. The main results are summarized in Table 1.

	Baseline	Separated	No Difference
Easier to locate target	5 (10.9%)	**33 (71.7%)**	8 (17.4%)
Sound Quality	10 (21.3%)	**33 (70.2%)**	4 (8.5%)
Better Background	19 (40.4%)	19 (40.4%)	9 (19.1%)

Table 1: User Survey Results, showing the number of responses for each question, out of 47 total respondents.

We see that most users found it easier to locate the target audio via the separated method, with only 5 choosing the baseline. There are likely two main effects at play here. The first is that by removing the target from the signals played from the microphone locations and concentrating it at one point, the spatial cues are much more consistent. The other is that because we have access to a source signal and its location, we are able to create a point source at that location. In the virtual speaker implementation, even if the source happens to be near a microphone and the signal is very dominant in that microphone, the spatial cues would be weaker because each microphone is rendered as a large-area audio source. One could improve this situation to some extent by reducing the radius of the virtual speaker audio sources, but the background becomes less diffuse for the microphones that do not carry the target signal as well.

It is encouraging that the overall subjective sound quality

was reported as generally better with the separated approach as well. Because sound quality is an ambiguous metric that means different things to different people it is difficult to interpret this precisely. From the free text responses it appears that people perceived the delayed echoes present in the baseline as unnatural-sounding, and others described the separated condition as "more natural". One user did report some "phasey" artifacts in the separated condition, and another compared it to wearing noise-canceling headphones. The general trends seem to be that the spatial distribution of the sound is more natural in the separated condition, but some work remains to ensure that the "holes" left by removing the target from the background don't degrade the background sound by leaving those areas of the spectrum with less energy. This is also consistent with the split opinions on the background sound.

In the free response section of the survey several users noted that the baseline background was somewhat louder than in the separated condition, which is likely due to energy removed in the filtering process. However, louder audio is generally correlated with better perceived quality, so the effect (if any) would be to bias in favor of the baseline.

It is important to note that this study is a relatively informal validation that we are on the right track. Because the participants were not blinded, their awareness of which method was which could bias their reported experience. This was primarily due to lack of time to perform a formal study, and the desire for the application to serve a dual purpose as a demonstration of the system.

Part V

Contributions and Conclusion

Contributions

This research spans several domains, and this document contains a mixture of introductions to existing techniques, descriptions of methods we've used, and the quantitative and qualitative results of our work. In this chapter we provide a high-level overview of the main contributions and lessons we've learned, that we hope will be of use to researchers and practitioners in the field.

Outdoor Audio Deployment

There have been many lessons we've learned deploying and maintaining an outdoor audio deployment, in **a distributed array unique in its duration of (almost) continuous operation, if not its overall scale**. The first, and perhaps most mundane, is that **cat5e cable makes a simple and effective four-channel audio cable**. The electrical specifications are similar to standard audio cables used in professional contexts, and there exists a wide variety of tooling developed for the telecommunications industry that eases installation considerably. In particular the use of a time-delay reflectometer is highly recommended for troubleshooting cable faults in the field. Animal damage is a common issue, and cables should be buried below ground.

Replacing cables with wireless communication seem like an obvious path to explore, but brings with it more challenges, particularly power, bandwidth, and synchronization. As demonstrated in this work, clock skew can be a substantial issue, and estimating the skew for each microphone individually adds an additional source of potential error.

Acoustic Measurement

Impulse response measurement (and more generally system identification) have been a widely-studied area of research. However, comparatively little work has been done in the audio community

to handle high noise environments, and accommodating deviations from the ideal linear, time-invariant system. This work highlights the fact that **transmitter/receiver synchronization is an important consideration for long-duration stimuli**. We have demonstrated that for the long-duration signals necessary to achieve acceptable signal-to-noise ratios, clock skew significantly degrades the measured impulse responses. In response, this work has also introduced **a new scheme for detecting and correcting clock skew**, and validated the algorithm in a real-world application.

Despite the popularity in the literature of a wide variety of impulse response measurement techniques, it is important to remember that **simple mechanical impulses can be effective for measuring basic acoustic propagation**, despite their relatively low SNR. They are also not subject to distortions caused by time-variance in the system. In addition to the time variance caused by clock skew, our results suggest that **environmental time-variance could be a significant issue in outdoor IR measurement**. Investigating this in more detail and characterizing the impact on different measurement techniques would require further study. We have also produced **a dataset of multichannel outdoor impulse responses from known locations**, which could be of use to acoustics researchers.

Localization

Many ideas and methods from existing literature can be re-evaluated and modified in the context of a large-scale array. One valuable outcome is the establishment of **SDR thresholds for TDOA localization based on cross-correlation**. This can help determine when these techniques would be expected to work well, and also provide a target for pre-localization noise reduction. This document also provides **a description of correlation bridging the statistical, linear algebraic, and signal processing perspectives**. While not novel research, bringing these perspectives together provides a deeper understanding and facilitates the application of techniques between domains and communities.

Signal Enhancement

The scale of our microphone array and the delays involved bring several novel challenges that are unusual in the field of multichannel signal enhancement. Prior work typically focuses on farfield

sources and small interchannel delays, and with array geometry known to a high precision (through design or calibration). This research proposes and validates **a signal enhancement algorithm based on rank-1 spatial covariance matrix factorization**, and demonstrates that this family of approaches can be applied in a large array context by first aligning the observed signals based on the source's location.

Soundscape Analysis and Resynthesis

The motivating application for this research is the ability to capture a spatially-varying soundscape and resynthesize it for a listener, giving them the ability to experience it from an arbitrary vantage point. Towards this end we have introduced **a framework for capturing, analyzing, and resynthesizing a soundfield**. We have presented **an end-to-end implementation that resynthesizes recordings from the field site**. Though the current implementation can only localize sources in the plane of the array, and the localization and foreground/background separation are limited to a single foreground source, these limitations are not fundamental to the framework and there are clear paths to overcoming them. Additionally, over the course of this research we have developed **a dataset of multichannel recordings with a variety of sounds played at known locations**. This dataset could be useful ground truth for source separation research in an outdoor large-array context.

Conclusion

This research proposes more questions than it answers, tilling fertile ground (sometimes literally) for a future research direction. Most signal processing and acoustics research uses simulations or audio recorded under controlled conditions, particularly in the multichannel domain. Working in the field gives new insights and exposes false assumptions. There are a number of improvements to make, open questions to answer, and follow-on studies to perform, but here I would like to chart out a larger-scale vision for the future of this work.

The current implementation of soundfield resynthesis has been shown to be an improvement over our previously-used baseline in a remote-presence application. Additionally individual improvements in latency, fidelity, and multi-source capability are all well within reach. With these improvements it will be feasible to integrate this resynthesis in on-site auditory augmented reality applications. These will provide transparent sensory augmentation (i.e. "super hearing") to enrich the listener's experience and connect them more deeply to the natural world around them.

In this dissertation I've described and validated the main building blocks necessary for capturing and analyzing a real site, as well as reproducing it for a listener, retaining their ability to move about the site freely and experience the soundscape from different points of view. As we live more and more of our lives in online and digital spaces, it is important to consider the sensory environment in which we're immersing ourselves. As the technology to render convincing and detailed spatial soundscapes continues to develop, we need frameworks and methods to capture those spaces. This creates the opportunity for telepresence to connect people to the real world, not just purely-virtual simulations and video games.

By situating this work in a wetland, I do not intend to replace a connection with nature, but rather to deepen it. One of the most powerful vignettes from Gershon Dublon's field experiments was with a user who found themselves really attending to the

soundscape of Tidmarsh for the first time, through the use of an unobtrusive piece of thoughtfully-designed technology. Understanding nature is a form of literacy and takes practice. The perception of an experienced birder or entomologist walking through the forest is qualitatively different from someone who spends their life behind a screen. I hope that this work can contribute to a richer sensory experience that brings attention to the diversity and abundance that might otherwise go unseen.

Glossary

Direction of Arrival (DoA) A spatial vector pointing from a source to the center of a sensor array. Generally measured in degrees or radians of azimuth (for 2D DoA) or azimuth and elevation (for 3D).

Farfield Given a sensor array, a source is in the farfield when it is far from the array relative to the distances between the sensors. In this regime arriving wavefronts can be thought of as plane waves, and time delays between microphones provide information only about the direction to the source, not its distance.

Nearfield Given a sensor array, a source is in the nearfield when the distance from the source to the sensors is similar to the distances between the sensors. The wavefront emanating from the source is spherical (assuming the medium is isotropic), and time delays provide information about source location.

Time Difference of Arrival (TDoA) A localization framework using the delay between when a target signal arrives at one sensor relative to another. It does not require knowledge of when the signal was emitted from the source.

Source Separation Extracting multiple signals of interest from a mixture, which might or might not include noise as well

Signal Enhancement Extracting a single target signal from noise.

Steered Response Power (SRP) A spatial function (of location or direction) that indicates how much power would be present in the output if the array were focused (in the beamforming sense) at that location or direction.

Beamforming A signal enhancement technique where noise is reduced by focusing the array at a particular location or direction. Focusing in this case generally means delaying and summing the individual array element signals to be consistent with that location and the array geometry.

Target The signal of interest, generally within a mixture

Interferer A non-noise signal that is not the target

Spatial Likelihood Function A spatial function used in localization (similar to the Steered Response Power) that gives the likelihood of the target being at that location or direction, under a particular model.

Interference Energy in a mixture signal that may have strong correlations or that we may have a model for, but that is not the target signal.

Noise Energy in a mixture signal that is not due to the target, and generally considered to be only weakly correlated in time, frequency, or space. More generally energy about which few modeling assumptions are made.

Image The component of the observed mixture signal that is due to the target. It often includes system effects such as reverberation, and in a multichannel case the image is multichannel. Source separation and signal enhancement are sometimes framed in terms of estimating the image of the source, rather than the source itself.

Spatialization The process of taking a source signal (generally single-channel) and performing simulation to compute the multichannel signal to present to a listener.

Localization Estimating the position of a sound source based on audio received at the microphones (or ears).

Stimulus A signal that is injected into a system to measure some properties of that system

Stimulus Response The signal observed from a system in response to some stimulus

Impulse Response The signal observed from a system in response to an impulse. Rather than measuring this directly it is often estimated from a stimulus response.

Time Aliasing Algorithmic noise that is introduced when frequency-domain processing is performed with insufficient resolution, which causes energy from the end of a signal to "wrap around" in time to the beginning.

Frequency Aliasing Algorithmic noise that is introduced when time- or space-domain processing is performed with insufficient

resolution, which causes energy from the positive frequencies of a signal to "wrap around" in frequency to the negative. For signals real-valued in time, the spectrum is conjugate symmetric and often only the positive frequencies are considered. In this case the frequency aliasing manifests as the high frequencies being reflected across the Nyquist frequency. This is often just called "Aliasing".

Spatial Aliasing Another name for Frequency Aliasing in the context where a signal is being measured at different points in space, but with insufficient resolution to uniquely determine the frequency.

Power Spectral Density (PSD) The expected power density of a signal as a function of frequency. That is, a measure of how much power a signal is expected to have near a given frequency. It is given by the Fourier transform of the signal's autocorrelation function, and is real-valued.

Cross Spectral Density (CSD) The frequency-domain correlation between two signals.

Musical Noise A type of noise often associated with spectral subtraction or other time-frequency processing that involves a threshold below which energy is removed. The noise is created by small "islands" in the time-frequency plane that are above the threshold, creating the percept of chirps or warbles (sometimes described as "watery" or "phasey" noise).

Wide-sense Stationary A signal whose mean and autocorrelation function are not a function of time. This is an important and widely-used signal model because it implies we can estimate the autocorrelation function by taking inner product between the signal and differently-delayed versions of itself.

Random process A random process is a distribution over signals - i.e. sampling from the process produces a signal.

Appendix A: User Survey

Users of the Soundscape Resynthesis Demo were asked to fill in a survey with the following questions:

Do you give permission for your responses to be used in Spencer Russell's dissertation and other published papers?

- Yes
- No

How much experience do you have working with sound?

- None
- Audio Hobbyist
- Audio Professional

How many audio samples did you listen to?

- 1
- 2-4
- 5-10
- 11-20
- More than 20

How much did you move around the environment?

- I stayed mostly where I started
- I explored

Did you hear a difference between the Baseline and Separated methods?

- Yes
- No

Which method had better sound quality?

- Baseline

- Separated
- No Difference

Which method made it easier to locate the target source (red orb)?

- Baseline
- Separated
- No Difference

Which method provided better background sound? (wind, etc., all the sound that's not the target).

- Baseline
- Separated
- No Difference

Please share any additional thoughts about your experience and expand on your answers above.

Free Text

References

Aarabi, Parham (2003). "The Fusion of Distributed Microphone Arrays for Sound Localization". In: *EURASIP Journal on Applied Signal Processing* 2003, pp. 338–347.

alienistcog (2014). *241974__alienistcog__2014-tree-frogs3.aiff*. URL: https://freesound.org/people/alienistcog/sounds/241974/.

Ananthabhotla, Ishwarya, David B. Ramsay, and Joseph A. Paradiso (2019). "HCU400: An Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 920–924.

Benaroya, Laurent et al. (2003). "Non Negative Sparse Representation for Wiener Based Source Separation with a Single Sensor". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference On*. Vol. 6. IEEE, pp. VI–613.

Berdahl, Edgar J. and Julius O. Smith (June 2008). *Transfer Function Measurement Toolbox*. URL: https://ccrma.stanford.edu/realsimple/imp_meas/Golay_Code_Theory.html (visited on 06/28/2019).

Berouti, Michael, Richard Schwartz, and John Makhoul (1979). "Enhancement of Speech Corrupted by Acoustic Noise". In: *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.* Vol. 4. IEEE, pp. 208–211.

Boll, Steven (1979). "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". In: *IEEE Transactions on acoustics, speech, and signal processing* 27.2, pp. 113–120.

Bryan, Nicholas J., Miriam A. Kolar, and Jonathan S. Abel (2010). "Impulse Response Measurements in the Presence of Clock Drift". In: *Audio Engineering Society Convention 129*. Audio Engineering Society.

Cano, E. et al. (Jan. 2019). "Musical Source Separation: An Introduction". In: *IEEE Signal Processing Magazine* 36.1, pp. 31–40. ISSN: 1053-5888. DOI: 10.1109/MSP.2018.2874719.

Cobos, M., A. Marti, and J. J. Lopez (Jan. 2011). "A Modified SRP-PHAT Functional for Robust Real-Time Sound Source Localization With Scalable Spatial Sampling". In: *IEEE Signal Processing Letters* 18.1, pp. 71–74. ISSN: 1070-9908. DOI: 10.1109/LSP.2010.2091502.

Collier, Travis C., Alexander N. G. Kirschel, and Charles E. Taylor (July 2010). "Acoustic Localization of Antbirds in a Mexican Rainforest Using a Wireless Sensor Network". In: *The Journal of the Acoustical Society of America* 128.1, pp. 182–189. ISSN: 0001-4966. DOI: 10.1121/1.3425729. URL: http://asa.scitation.org/doi/10.1121/1.3425729 (visited on 04/09/2020).

Cornell Guide to Bird Sounds: Master Set for North America (2014). Ithaca, New York.

DiBiase, Joseph Hector (May 2000). "A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays". en. PhD. Providence, RI: Brown University.

Dublon, Gershon (2018). "Sensor(y) Landscapes:Technologies for New Perceptual Sensibilities". Doctoral Dissertation. Cambridge, MA: Massachusetts Institute of Technology.

Dublon, Gershon and Joseph A. Paradiso (July 2014). "Extra Sensory Perception: How a World Filled with Sensors Will Change the Way We See, Hear, Think and Live". en. In: *Scientific American* 311.1. DOI: 10.1038/scientificamerican0714-36. URL: https://www.scientificamerican.com/article/how-a-sensor-filled-world-will-change-human-consciousness/ (visited on 04/09/2020).

Duhart, Clement et al. (Oct. 2019). "Deep Learning for Environmental Sensing Toward Social Wildlife Database". In: *The Ninth International Workshop on Climate Informatics*. Paris, France.

Duong, Ngoc QK, Emmanuel Vincent, and Rémi Gribonval (2010). "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.7, pp. 1830–1840. URL: http://ieeexplore.ieee.org/abstract/document/5466223/.

Farina, Angelo (2000). "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique". In: *Audio Engineering Society Convention 108*. Audio Engineering Society.

Fontaine, Mathieu et al. (2017). "Explaining the Parameterized Wiener Filter with Alpha-Stable Processes". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. URL: https://hal.archives-ouvertes.fr/hal-01548508/ (visited on 09/22/2017).

Foster, S. (Apr. 1986). "Impulse Response Measurement Using Golay Codes". In: *ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 11, pp. 929–932. DOI: 10.1109/ICASSP.1986.1168980.

FunkApache (2017). *393389__funkapache__cricket.wav*. URL: https://freesound.org/people/FunkApache/sounds/393389/.

Gallager, Robert G (2008). *Circularly-Symmetric Gaussian Random Vectors*. en. Tech. rep., p. 9.

Gamper, Hannes (2017). "Clock Drift Estimation and Compensation for Asynchronous Impulse Response Measurements". en. In: *2017 Hands-Free Speech Communications and Microphone Arrays (HSCMA)*. San Francisco, CA, USA: IEEE, pp. 186–190. ISBN: 978-1-5090-5925-6. DOI: 10.1109/HSCMA.2017.7895587. URL: http://ieeexplore.ieee.org/document/7895587/ (visited on 09/23/2019).

Golay, M. (Apr. 1961). "Complementary Series". In: *IRE Transactions on Information Theory* 7.2, pp. 82–87. ISSN: 0096-1000. DOI: 10.1109/TIT.1961.1057620.

Gorzel, Marcin et al. (Mar. 2019). "Efficient Encoding and Decoding of Binaural Sound with Resonance Audio". English. In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=20446 (visited on 04/02/2020).

Guggenberger, Mario, Mathias Lux, and Laszlo Böszörmenyi (2015). "An Analysis of Time Drift in Hand-Held Recording Devices". In: *International Conference on Multimedia Modeling*. Springer, pp. 203–213.

Haddad, Don Derek et al. (2017). "Resynthesizing Reality: Driving Vivid Virtual Environments from Sensor Networks". en. In: ACM Press, pp. 1–2. ISBN: 978-1-4503-5008-2. DOI: 10.1145/3084363.3085027. URL: http://dl.acm.org/citation.cfm?doid=3084363.3085027 (visited on 06/25/2018).

Hasegawa, Keisuke et al. (2010). "Blind Estimation of Locations and Time Offsets for Distributed Recording Devices". en. In: *Latent Variable Analysis and Signal Separation*. Ed. by Vincent Vigneron et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 57–64. ISBN: 978-3-642-15995-4. DOI: 10.1007/978-3-642-15995-4_8.

Hennequin, Romain et al. (2019). "Spleeter: A Fast and State-of-the Art Music Source Separation Tool with Pre-Trained Models". en. In: p. 2.

Holters, Martin, Tobias Corbach, and Udo Zölzer (2009). "Impulse Response Measurement Techniques and Their Applicability

in the Real World". en. In: *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*. Milan, Italy, p. 5.

Ito, Keith (2017). *The LJ Speech Dataset*. URL: https://keithito.com/LJ-Speech-Dataset/.

Jacovitti, Giovanni and Gaetano Scarano (1993). "Discrete Time Techniques for Time Delay Estimation". In: *IEEE Transactions on signal processing* 41.2, pp. 525–533.

Jourjine, Alexander, Scott Rickard, and Ozgur Yilmaz (2000). "Blind Separation of Disjoint Orthogonal Signals: Demixing N Sources from 2 Mixtures". In: *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference On*. Vol. 5. IEEE, pp. 2985–2988.

Kahl, Stefan et al. (2019). "Overview of BirdCLEF 2019: Large-Scale Bird Recognition in Soundscapes". en. In: p. 9.

kayceemixer (2014). *251495__kayceemixer__kc-animal-frog-lakeside-penticton-2013.wav*. URL: https://freesound.org/people/kayceemixer/sounds/251495/.

Knapp, C. and G. Carter (Aug. 1976). "The Generalized Correlation Method for Estimation of Time Delay". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24.4, pp. 320–327. ISSN: 0096-3518. DOI: 10.1109/TASSP.1976.1162830.

Le Roux, Jonathan and Emmanuel Vincent (Mar. 2013). "Consistent Wiener Filtering for Audio Source Separation". In: *IEEE Signal Processing Letters* 20.3, pp. 217–220. ISSN: 1070-9908. DOI: 10.1109/LSP.2012.2225617.

Loizou, Philipos C. (2007). "Wiener Filtering". In: *Speech Enhancement: Theory and Practice*. 1st, pp. 143–212. URL: https://www.crcpress.com/Speech-Enhancement-Theory-and-Practice-Second-Edition/Loizou/p/book/9781138075573 (visited on 09/19/2017).

Lombardi, Michael A. (2010). "Time and Frequency from A to Z". In: URL: https://www.nist.gov/pml/time-and-frequency-division/popular-links/time-frequency-z/time-and-frequency-z-f.

Lynch, Evan F and Joseph A Paradiso (2016). "SensorChimes: Musical Mapping for Sensor Networks". en. In: *New Interfaces for Musical Expression 2016*. Brisbane, Australia, p. 6.

Mandel, Michael I., Ron J. Weiss, and Daniel PW Ellis (2010). "Model-Based Expectation-Maximization Source Separation and Localization". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.2, pp. 382–394. URL: http://ieeexplore.ieee.org/abstract/document/5200357/.

Martin, Rainer (1994). "Spectral Subtraction Based on Minimum Statistics". In: *power* 6.8.

Mateljan, I. (1999). "Signal Selection for the Room Acoustics Measurement". In: *1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 199–202. DOI: 10.1109/ASPAA.1999.810884.

Mayton, Brian et al. (May 2017). "The Networked Sensory Landscape: Capturing and Experiencing Ecological Change Across Scales". In: *Presence: Teleoperators and Virtual Environments* 26.2, pp. 182–209. ISSN: 1054-7460. DOI: 10.1162/PRES_a_00292. URL: https://doi.org/10.1162/PRES_a_00292 (visited on 06/06/2018).

Mennill, Daniel J. et al. (Apr. 2006). "Accuracy of an Acoustic Location System for Monitoring the Position of Duetting Songbirds in Tropical Forest". In: *The Journal of the Acoustical Society of America* 119.5, pp. 2832–2839. ISSN: 0001-4966. DOI: 10.1121/1.2184988. URL: https://asa-scitation-org.libproxy.mit.edu/doi/abs/10.1121/1.2184988 (visited on 06/14/2019).

Mills, D. (Mar. 1992). *Network Time Protocol (Version 3) Specification, Implementation and Analysis*. en. Tech. rep. RFC1305. RFC Editor. DOI: 10.17487/rfc1305. URL: https://www.rfc-editor.org/info/rfc1305 (visited on 03/24/2020).

Oppenheim, Alan V. and George C. Verghese (2015). "Wiener Filtering". In: *Signals, Systems and Inference*. Pearson.

– (2010). *Signals, Systems, and Inference: Class Notes for 6.011: Introduction to Communication, Control and Signal Processing Spring 2010*. en.

Ozerov, Alexey and Cédric Févotte (2009). "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation". In: *IEEE Transactions on Audio, Speech, and Language Processing* 18.3, pp. 550–563.

Raffel, Colin et al. (2014). "MIR_EVAL: A Transparent Implementation of Common MIR Metrics". In: *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*. Ed. by Hsin-Min Wang, Yi-Hsuan Yang, and Jin Ha Lee, pp. 367–372. URL: http://www.terasoft.com.tw/conf/ismir2014/proceedings/T066_320_Paper.pdf (visited on 02/15/2016).

Rockah, Y. and P. Schultheiss (June 1987). "Array Shape Calibration Using Sources in Unknown Locations–Part II: Near-Field Sources and Estimator Implementation". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 35.6, pp. 724–735. ISSN: 0096-3518. DOI: 10.1109/TASSP.1987.1165222.

Roux, Jonathan Le et al. (Nov. 2018). "SDR - Half-Baked or Well Done?" en. In: *arXiv:1811.02508 [cs, eess]*. arXiv: 1811.02508 [cs,

eess]. URL: http://arxiv.org/abs/1811.02508 (visited on 01/20/2020).

Russell, Spencer, Gershon Dublon, and Joseph A. Paradiso (2016). "HearThere: Networked Sensory Prosthetics Through Auditory Augmented Reality". en. In: *Augmented Human*. ACM Press, pp. 1–8. ISBN: 978-1-4503-3680-2. DOI: 10.1145/2875194.2875247. URL: http://dl.acm.org/citation.cfm?doid=2875194.2875247 (visited on 06/25/2018).

Schafer, R. Murray (1977). *The Tuning of the World*. Alfred A. Knopf.

Schmidt, R. (Mar. 1986). "Multiple Emitter Location and Signal Parameter Estimation". In: *IEEE Transactions on Antennas and Propagation* 34.3, pp. 276–280. ISSN: 0018-926X. DOI: 10.1109/TAP.1986.1143830.

Stahnke, Wayne (1973). "Primitive Binary Polynomials". en-US. In: *Mathematics of Computation* 27.124, pp. 977–980. ISSN: 0025-5718, 1088-6842. DOI: 10.1090/S0025-5718-1973-0327722-7. URL: http://www.ams.org/home/page/ (visited on 07/18/2018).

Sumitani, S. et al. (May 2019). "An Integrated Framework for Field Recording, Localization, Classification and Annotation of Birdsongs Using Robot Audition Techniques — Harkbird 2.0". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8246–8250. DOI: 10.1109/ICASSP.2019.8683743.

Svensson, Peter and Johan L. Nielsen (May 1996). "Errors in MLS Measurements Caused by Time Variance in Acoustic Systems". English. In: *Audio Engineering Society Convention 100*. Audio Engineering Society. URL: http://www.aes.org/e-lib/browse.cfm?elib=7509 (visited on 09/23/2019).

Traer, James and Josh H. McDermott (Nov. 2016). "Statistics of Natural Reverberation Enable Perceptual Separation of Sound and Space". en. In: *Proceedings of the National Academy of Sciences* 113.48, E7856–E7865. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1612524113. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1612524113 (visited on 10/02/2017).

Valenzise, Giuseppe et al. (2007). "Scream and Gunshot Detection and Localization for Audio-Surveillance Systems". In: *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference On*. IEEE, pp. 21–26.

Valle, Andrea, Vincenzo Lombardo, and Mattia Schirosa (2009). "Simulating the Soundscape through an Analysis/Resynthesis Methodology". en. In: *Auditory Display*. Vol. 5954. Copenhagen, Denmark: Springer, pp. 330–357. ISBN: 978-3-642-12438-9 978-3-642-12439-6. DOI: 10.1007/978-3-642-12439-6_17. URL:

http://link.springer.com/10.1007/978-3-642-12439-6_17 (visited on 04/09/2020).

Wang, Zhong-Qiu, Xueliang Zhang, and DeLiang Wang (Sept. 2018). "Robust TDOA Estimation Based on Time-Frequency Masking and Deep Neural Networks". en. In: *Interspeech 2018*. ISCA, pp. 322–326. DOI: 10.21437/Interspeech.2018-1652. URL: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/1652.html (visited on 06/20/2019).

Ward, Dominic et al. (2018). "SiSEC 2018: State of the Art in Musical Audio Source Separation - Subjective Selection of the Best Algorithm". en. In: p. 4.

Wijers, Matthew et al. (Nov. 2019). "CARACAL: A Versatile Passive Acoustic Monitoring Tool for Wildlife Research and Conservation". In: *Bioacoustics* 0.0, pp. 1–17. ISSN: 0952-4622. DOI: 10.1080/09524622.2019.1685408. URL: https://doi.org/10.1080/09524622.2019.1685408 (visited on 04/11/2020).

Yilmaz, Ozgur and Scott Rickard (2004). "Blind Separation of Speech Mixtures via Time-Frequency Masking". In: *IEEE Transactions on signal processing* 52.7, pp. 1830–1847.

Zahorik, Pavel (Feb. 2000). "Limitations in Using Golay Codes for Head-Related Transfer Function Measurement". In: *The Journal of the Acoustical Society of America* 107.3, pp. 1793–1796. ISSN: 0001-4966. DOI: 10.1121/1.428579. URL: https://asa.scitation.org/doi/abs/10.1121/1.428579 (visited on 12/11/2018).